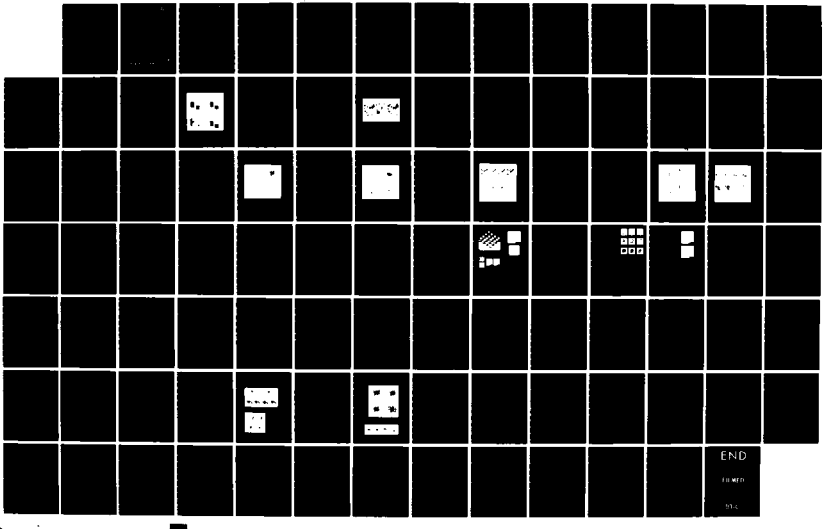
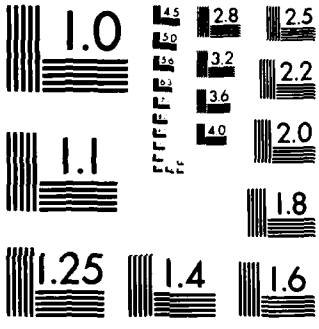
AD-A149 344 DIFFRACTION-LIMITED IMAGING OF SPACE OBJECTS II(U) 1/1
ENVIRONMENTAL RESEARCH INST OF MICHIGAN ANN ARBOR
J R FIENUP OCT 84 ERIM-161900-13-T AFOSR-TR-84-1171
UNCLASSIFIED F49620-82-K-0018 F/G 17/8 NL





Annual ~~Interim~~ Scientific Report

DIFFRACTION-LIMITED IMAGING OF SPACE OBJECTS II

1 March 1983 through 29 February 1984

J.R. FIENUP
Infrared and Optics Division

OCTOBER 1984

Director, Physical and Geophysical Sciences
Air Force Office of Scientific Research/NP
Building 410, Bolling AFB, D.C. 20332
Contract No.: F49620-82-K-0018

ENVIRONMENTAL
RESEARCH INSTITUTE OF MICHIGAN
BOX 8618 • ANN ARBOR • MICHIGAN 48107

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A149 344

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 161900-13-T	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 84-1171
6a. NAME OF PERFORMING ORGANIZATION Environmental Research Institute of Michigan 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research/NP
6c. ADDRESS (City, State and ZIP Code) P.O. Box 8618 Ann Arbor, MI 48107	7b. ADDRESS (City, State and ZIP Code) Building 410 Bolling AFB, DC 20332
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR 8b. OFFICE SYMBOL (If applicable) NP	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-82-K-0018
8c. ADDRESS (City, State and ZIP Code) Bldg 410 Bolling AFB 20332-6448	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO.: 61102F; PROJECT NO.: 2311; TASK NO: A1; WORK UNIT NO:
11. TITLE (Include Security Classification) Diffraction-Limited Imaging of Space Objects II	
12. PERSONAL AUTHOR(S) Fienup, James R.	
13a. TYPE OF REPORT Annual/Scientific Rept 13b. TIME COVERED FROM 1 Mar 83 TO 29 Feb 84	14. DATE OF REPORT (Yr., Mo., Day) 1984, October 15. PAGE COUNT ix+86
16. SUPPLEMENTARY NOTATION The AFOSR project scientist was Dr. Henry Radoski	

17 COSATI CODES		
FIELD	GROUP	SUB. GR.
20	06	
03		

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Phase retrieval, Space object imaging, Reconstruction algorithms, Image reconstruction, Astronomical speckle interferometry

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

→This report describes the results of the second year of a three-year research program to investigate methods for obtaining diffraction-limited images of space objects, despite the turbulent atmosphere, by reconstructing images from data provided by optical interferometers (particularly stellar speckle interferometry). Accomplishments include the following. (1) Improved image reconstruction algorithms were developed. (2) A better understanding of modes of stagnation of algorithms was developed. (3) The performance of the shift-and-add image formation method and of one recursive algorithm were investigated. (4) A second recursive algorithm was shown to suffer from a uniqueness problem. (5) A potential new remote sensing application of the iterative reconstruction algorithm was explored.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Henry Radoski	22b. TELEPHONE NUMBER (Include Area Code) (202)767-4906 22c. OFFICE SYMBOL NP

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

FOREWORD

This report was prepared by the Infrared and Optics Division of the Environmental Research Institute of Michigan. The work was sponsored by the Air Force Office of Scientific Research/AFSC, United States Air Force, under Contract No. F49620-82-K-0018.

This interim scientific report covers work performed between 1 March 1983 and 29 February 1984. The contract monitor is Dr. Henry Radoski, Directorate of Physical and Geophysical Sciences, AFOSR/NP, Building 410, Bolling Air Force Base, D.C. 20332. The principal investigator is James R. Fienup. Major contributors to the effort are James R. Fienup and Christopher C. Wackerman. Additional contributors to the effort are Thomas R. Crimmins and Susan C. Elm.

SUMMARY

This report describes the results of the second year of a three-year research program to investigate methods for obtaining diffraction-limited images of space objects, despite the turbulent atmosphere, by reconstructing images from data provided by optical interferometers (particularly stellar speckle interferometry). Accomplishments include the following. (1) Improved image reconstruction algorithms were developed. (2) A better understanding of modes of stagnation of algorithms was developed. (3) The performances of the shift-and-add image formation method and of one recursive algorithm were investigated. (4) A second recursive algorithm was shown to suffer from a uniqueness problem. (5) A potential new remote sensing application of the iterative reconstruction algorithms was explored.

CONTENTS

LIST OF FIGURES

DIFFRACTION-LIMITED IMAGING OF SPACE OBJECTS II
1 March 1983 to 29 February 1984

1
INTRODUCTION AND OBJECTIVES

This report describes the results of the second year's effort in a three-year research program to investigate methods of obtaining diffraction-limited images of space objects, despite the turbulent atmosphere, by reconstructing images from data provided by optical interferometers (particularly stellar speckle interferometry).

Atmospheric turbulence typically limits the angular resolution of earth-bound optical telescopes to one second of arc or worse, which is *fifty times poorer than the theoretical diffraction* limit of a 5-meter optical telescope. It is possible to gather diffraction-limited information through the turbulent atmosphere by a variety of interferometric techniques, including Michelson stellar interferometry [1], intensity interferometry [2], amplitude interferometry [3], and stellar speckle interferometry [4, 5]. However, this diffraction-limited information is in the form of the modulus (magnitude) of the Fourier transform of the object being viewed. Until recently only the autocorrelation of the object, but not the object itself, could be reconstructed from this data, except for special cases.

In recent years an iterative method [6-9] has been developed for reconstructing an object from its Fourier modulus, thereby making possible the reconstruction of diffraction-limited imagery from interferometer data. The algorithm utilizes the measured Fourier modulus data as well as (1) the a priori information that the object's spatial (or angular) brightness distribution is a nonnegative function

and (2) information about the object's diameter which can be computed from the autocorrelation function. The algorithm and its numerous applications is described in detail in Appendix A [9].

The goal of the program is to further investigate and develop this method of obtaining diffraction-limited images. Included in the three-year program are investigations into improving the reconstruction algorithm, developing methods for processing noisy astronomical data, studying the uniqueness of the reconstruction, and investigating ways to increase the spectral bandwidth of stellar speckle interferometry. In the second year of the effort, the emphasis was on developing new and improved reconstruction algorithms. Initial studies of the uniqueness problem and of the properties of astronomical data were also begun.

The research accomplishments for the second year are summarized in Section 2 and are described in more detail in Sections 3 through 8 and in the Appendices. References are listed in Section 9.

2
RESEARCH ACCOMPLISHMENTS

The second year of research effort can be divided into seven major topics.

1. The new recursive algorithm [10] described in last year's report [11] was implemented and tested both on noise-free and on noisy Fourier modulus data.
2. Improvements in the iterative Fourier transform reconstruction algorithm [6-8] were made that enable one to reconstruct difficult objects that previously defied reconstruction attempts.
3. Alternative iterative algorithms were devised.
4. Investigations were made into the problem of stripes in the reconstructed images.
5. The shift-and-add algorithm was implemented and tested on a complicated extended object.
6. Results were obtained indicating a possible new remote sensing application of the iterative reconstruction algorithm.
7. Recently published claims regarding the uniqueness of phase retrieval when the edges of the object are known were shown to be false by counterexample [12].

Recent publications arising from this work are References 10 and 12-18. Reference 9 is noted as a recent related publication arising from a previous research program [19] and is included as Appendix A. References 10, 16, 17, 18 and 12 are included as Appendices B, C, D, E and F, respectively.

The seven topics are briefly described in the remainder of this section and are described in more detail in Sections 3 to 8 and in the Appendices.

2.1 NEW RECURSIVE ALGORITHM

As described in last year's report [11] and in Appendix B, a new recursive algorithm has been developed which is capable of reconstructing an object from its autocorrelation function, which can be computed from the modulus of its Fourier transform. It works for objects having latent reference points--unresolved points within the object field that are not sufficiently far from the main part of the object to satisfy the condition for holography, but satisfy weaker conditions. The recursive algorithm was coded on a computer and exercised on two different types of objects using autocorrelations having a variety of signal-to-noise ratios. As expected the recursive algorithm was fairly sensitive to noise, making it less practical for real-world applications than the iterative Fourier transform algorithm [6-8]. Improvements were made in the recursive algorithm to make it somewhat less sensitive to noise. A more detailed description of this work is given in Section 3.

2.2 IMPROVEMENTS IN THE ITERATIVE ALGORITHM

The iterative Fourier transform algorithm has been particularly successful on objects having complicated shapes, such as satellites [6, 7, 17]. However, for some types of objects, such as those whose support fills a square, the algorithm has a tendency to stagnate without finding a solution [20]. Two modifications of the basic approach were demonstrated for overcoming this problem. The first was to employ the defogging method of Bates and Fright [21]. The second modification is to break the symmetry of the partially reconstructed image in order to

allow the algorithm to converge to one of the two possible solutions. This is described further and an example is shown in Section 4.

2.3 ALTERNATIVE ITERATIVE ALGORITHM

The theoretical justification for the input-output iterative Fourier transform algorithm [6-9] alludes to a control theory point-of-view. Yet rigorous control theory had not actually been applied to the problem. Alternative algorithms based on control theory are presented in Section 5. Further elaboration of these algorithms and their implementation and testing will be required to determine whether they will offer improved performance over existing algorithms.

2.4 INVESTIGATION OF STRIPES

In some cases the output of the iterative algorithm has the appearance of original object but with a pattern of low-contrast stripes superimposed [22, 17]. The phenomenon of stripes appearing in the reconstructed image was extensively investigated. Properties of the phase of the Fourier transform of the striped image were studied. Several methods for removing the stripes were investigated. We feel that we are on the threshold of solving this problem, as described in Section 6. When this problem is completely solved, then it will be possible to answer the question of the uniqueness of the reconstructed image (see Appendix D) more definitively.

2.5 SHIFT-AND-ADD

The shift-and-add method of imaging from short-exposure astronomical images has in the past been exercised primarily for very simple objects having in their field-of-view very bright unresolved points [23, 24]. The shift-and-add method was attempted both on a

point-like object and on a more realistic object--a satellite having strong glints. Although the result for the point-like object was very good, the result for the extended object was very poor, indicating that the shift-and-add method is not appropriate for complicated extended objects. These results are shown in Section 7.

2.6 NEW ITERATIVE APPLICATION

A new remote sensing application for the iterative Fourier transform algorithm was developed under ERIM internal funding [25]. It permits the operation of, say, a synthetic aperture radar system having reduced performance requirements for the phase stability of its local oscillator and motion compensation. It might also be useful for the electron microscopy phase retrieval problem. It involves the iterative retrieval of phase using a single intensity measurement plus a shape constraint on the object or upon the pattern of radiation by which the object is illuminated. Under the present program the issue of the shape constraint was explored further. It was found that certain interesting shapes are sufficient for reconstructing a complex-valued object function from the magnitude of its Fourier transform. The reconstruction algorithm and some reconstruction results are shown in Section 8.

2.7 AMBIGUITY OF PHASE RETRIEVAL USING BOUNDARY CONDITIONS

Claims have been made that an object can be uniquely reconstructed from its Fourier modulus via the autocorrelation function if the values of the edges of the object are known [26]. It is shown in Appendix F that knowledge of the autocorrelation function and of the boundary values of the object are not sufficient to uniquely specify the object in all cases. It is further shown how and why the recursive algorithm of Hayes and Quatieri [26] fails for the nonunique cases. Recently it

has also been shown that the recursive algorithm [26] can fail even when the object is uniquely related to its Fourier modulus. An example of where it fails for a unique object is in the case of an object like that shown in Figure 1a of Appendix F but with the value of 4 in the second column from the right and the second line from the top replaced by any other value. This last result will be described in more detail in a later report.

3
EXPERIMENTAL RESULTS USING NEW RECURSIVE ALGORITHM

As described in Appendix B, a new recursive algorithm has been developed for reconstructing an object from its autocorrelation function, which can be computed from the modulus of the Fourier transform of the object. It is applicable to objects having latent reference points [10], and knowledge of the support of the object may be required. In this section examples of reconstruction experiments using the recursive algorithm are shown.

Figure 3-1 shows results of the recursive reconstruction algorithm for which the Fourier modulus (or autocorrelation) data was corrupted with varying amounts of noise. The object consists of an equilateral right triangle of 16 pixels on each side having a brighter rectangle and a brighter square imbedded in it. It is assumed known that the object's support is the triangle. Figure 3-1(a) shows the original object. Figure 3-1(b), (c) and (d) are the reconstructed images when the root-mean-squared (RMS) error of the Fourier modulus data was 0.005175, 0.05585 and 0.01795, respectively. The RMS error of these reconstructed images is 0.0400, 0.6088 and 0.1390, respectively. That is, for the noisiest case of Figure 3-1(c), a 5.6 percent error in the data resulted in a 60.9 percent error in the reconstructed image.

To get a feel for how bad a 60 percent error is, consider the following. Suppose the object were constant, equal to unity, over the known region of support (within the triangle). If the reconstructed image were a set of random numbers uniformly distributed between 0 and b, then the rms error for the optimum value of b can be shown to be 50 percent. That is, the reconstructed image shown in Figure 3-1(c), having RMS error of 60 percent, is worse than a reconstructed image

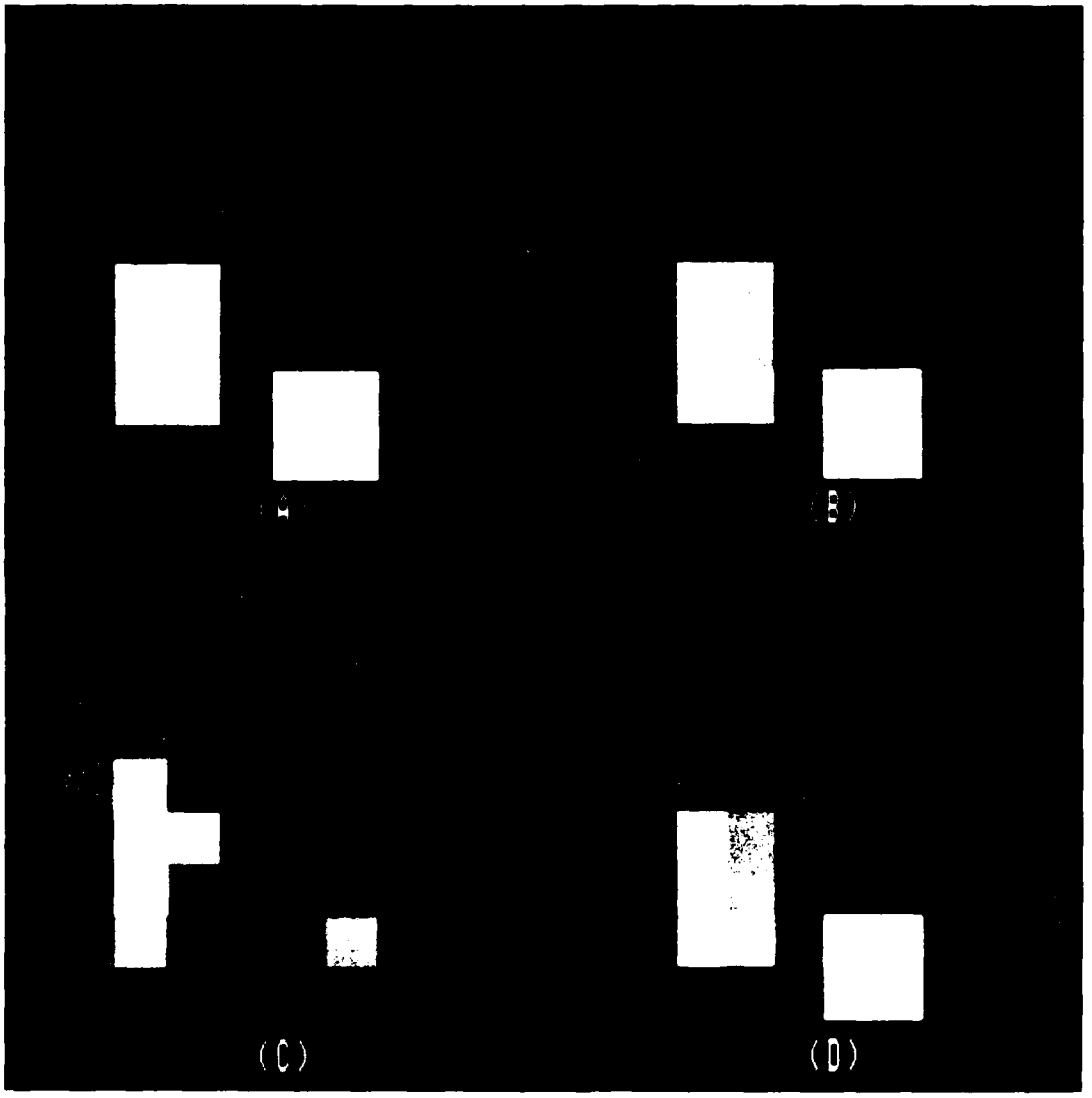


Figure 3-1. Images Reconstructed Using Recursive Algorithm. (a) Original object; (b)-(d) images reconstructed from autocorrelations having root-mean-squared error (b) 0.0400, (c) 0.6088, and (d) 0.01795.

consisting of random numbers.

By comparison, in experiments using the iterative Fourier transform algorithm (on a different object), a 5% error in the data resulted in a 20% error in the reconstructed image [22]. Therefore from this limited experience it appears that the recursive algorithms is, as predicted [10], highly sensitive to noise. The iterative Fourier transform algorithm would appear to be the preferred method of image reconstruction.

For the triangular support case, one can generate as many as three separate estimates for each value, one associated with each of the three corner pixels. In the absence of noise these three reconstructions are identical, but with noise present they will in general be different, and an improved algorithm would decrease noise effects by averaging the three estimates. This method was tried and the following was discovered. The computation of a value depends not only the corner pixel but also on a number of previously reconstructed values as well. In the presence of noise each of these reconstructed values will have some error associated with it, and the more of them that are used to reconstruct a new value, the more error that new value will have. It was found that the estimate that is generated from the maximum number of previously reconstructed values has accumulated so much error that including its value in the average degrades rather than improves the reconstruction. The optimum number of estimates to use was found to depend on the signal-to-noise ratio and on distance from the edges of the triangle. For most values, only one or two estimates was optimal. This resulted in a modest improvement over the algorithm employing only a single estimate.

4
IMPROVEMENTS IN THE ITERATIVE FOURIER TRANSFORM ALGORITHM

Although the iterative Fourier transform algorithm has been shown to be successful for space objects such as satellites [6, 7, 17], it can have problems converging for some other types of objects. For example, for the object shown in Figure 4-1(a), a picture of a bird bounded by a square, the algorithm has a strong tendency to stagnate without finding a solution [20]. Two methods were demonstrated for overcoming this problem and converging to a solution: the defogging method of Bates and Fright [21] and a new method of temporarily using a reduced-area asymmetric support constraint. The latter method seems to be the more important of the two.

The defogging method attempts to compensate for the fact that a low-contrast object on a bright background causes very little "interference," that is, its Fourier transform has not much structure. The defogging method consists of reducing the large central lobe of the Fourier modules, raising the relative values at the higher spatial frequencies. In the image domain this corresponds to reducing any slowly-varying bias-like (or fog) component of the image, thereby emphasizing the finer-structure details. Phase retrieval algorithms that work poorly on a low-contrast image tend to work better on the defogged version of the image. After the defogged image is reconstructed, the slowly-varying fog component is added back in. As a final step the refogged image is refined by further iterations of phase retrieval.

A reduced-area asymmetric support constraint is used for types of objects which often cause the iterative Fourier transform algorithm to stagnate on a partially reconstructed image. Recall that there is

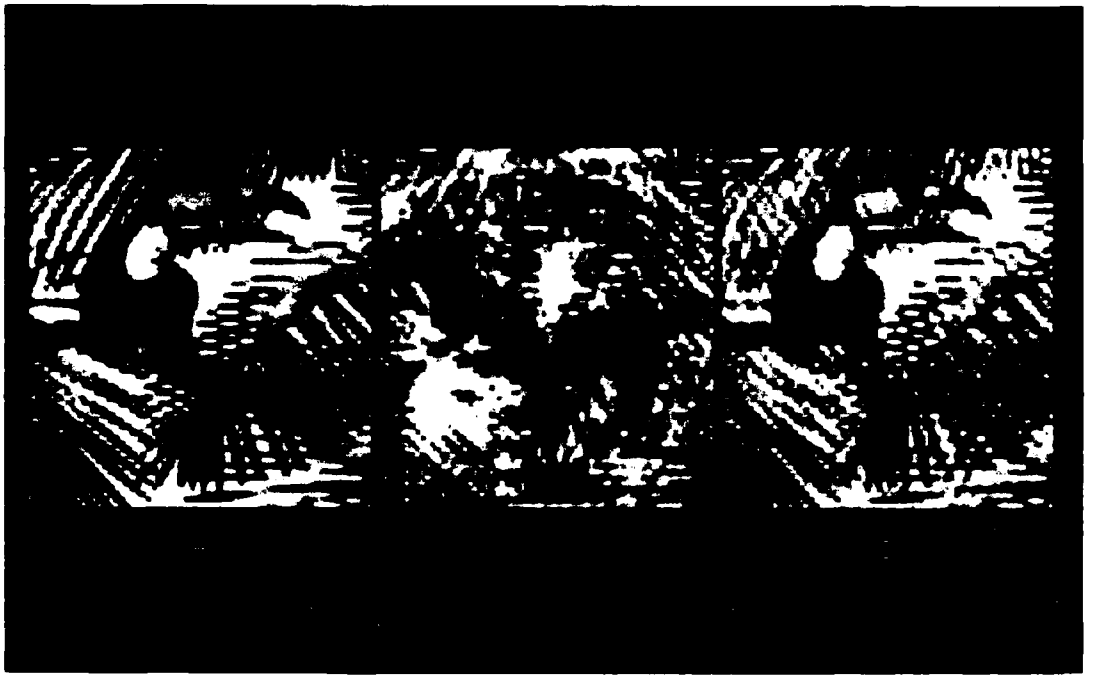


Figure 4-1. Use of Asymmetric Support Constraint. (a) Original object; (b) output image from iterative Fourier transform algorithm which has stagnated; (c) output image after application of reduced-area support constraint followed by further iterations.

always a two-fold ambiguity: f(-x, -y) and f(x, y) have the same Fourier modulus. The ambiguous image f(-x, -y) is just f(x, y) rotated by 180° which is equivalent to being reflected through the origin. When there is a symmetric support as in the case of the object shown in Figure 4-1(a), the algorithm may stagnate with an output such as the one shown in Figure 4-1(b), which has features of both the object and the 180° rotated object. The output image changes little with further iterations. Apparently the algorithm gets stuck when it is half-way between the two different solutions: it is unable to shake one off and converge to the other. In the case where an asymmetric support constraint is known, this particular mode of stagnation tends not to occur since the asymmetric support constraints moves the solution toward f(x, y) and away from f(-x, -y).

The method of using a reduced-area asymmetric support constraint is as follows. A support constraint is defined that includes one side (including edges) of the object but not the other side and edges. This support constraint is chosen to be smaller than the known support of the object and to be as asymmetric as possible, so that it has little in common with the 180°-rotated version of the support constraint. A few iterations are then performed with the reduced-area asymmetric support constraint (rather than using the correct support constraint). It is hoped that this causes one of the two images, f(x, y) or f(-x, -y) to be preferentially enhanced over the other. After switching back to the correct support constraint, either f(x, y) or f(-x, -y) will be strong enough compared with the other that upon further iterations the algorithm converges to the stronger one and away from the weaker one.

Figure 4-1(c) shows the reconstructed output image after using both the defogging method and the reduced-area asymmetric support constraint for a few iterations then continuing with further iterations using the correct support constraint and the original Fourier modulus data.

Comparing it with the output image shown in Figure 4-1(b), it is seen that these techniques yielded much better results in this case.

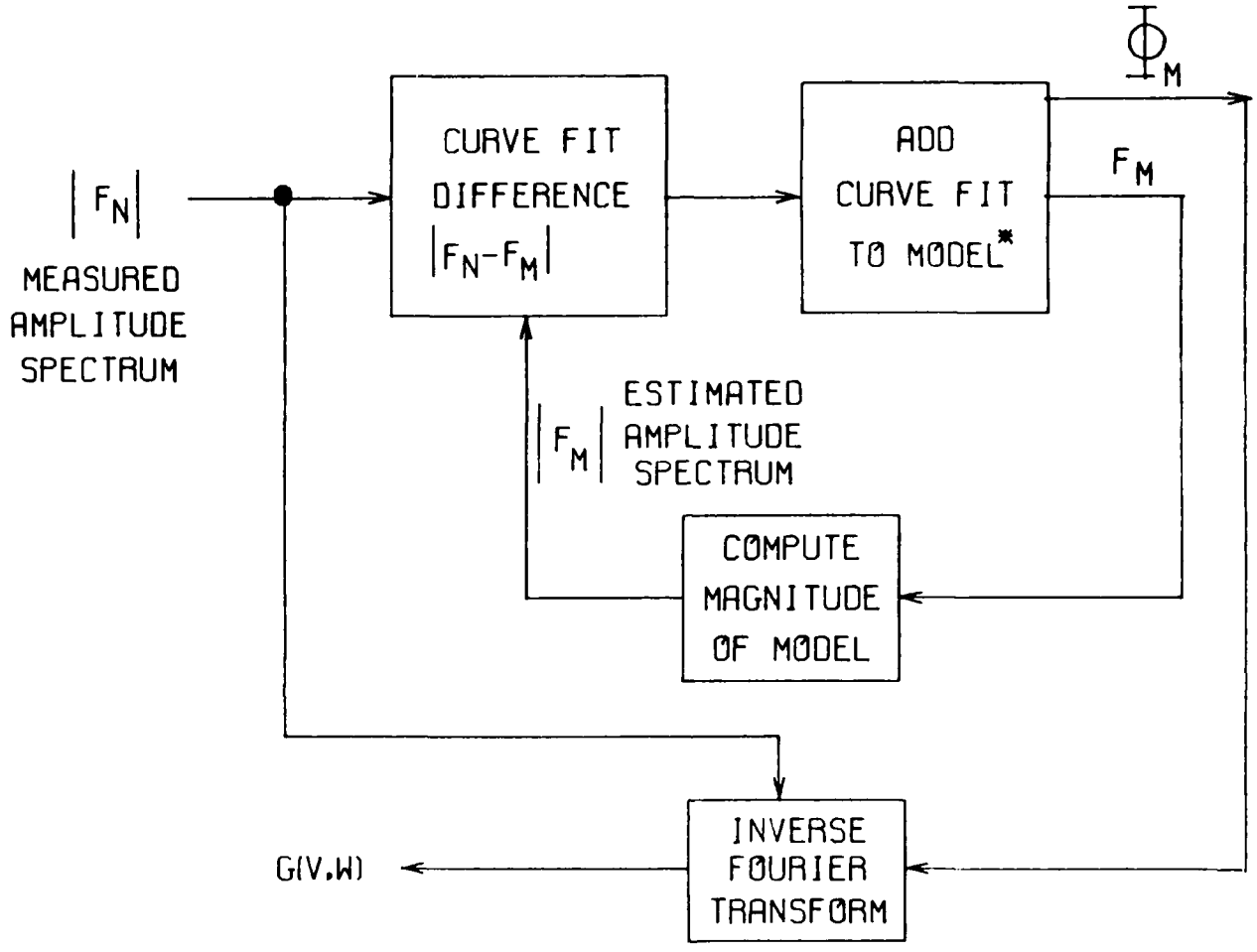
These methods have been exercised in only very limited circumstances and have not yet been optimized and automated. Further work to develop these promising methods is clearly called for. Such auxiliary procedures are not necessary for the objects that are easier to reconstruct but are very important for the more difficult cases.

5
ALTERNATIVE ITERATIVE ALGORITHMS

The iterative Fourier transform algorithm, which is described in detail in Appendix A, works very well in a wide range of situations but converges slowly or not at all in some cases. Furthermore it is always desired to arrive at a solution using fewer iterations and less computer time. For these reasons we are always looking for ways to improve the existing algorithms or devise alternative algorithms that converge faster. The two algorithms shown in Figures 5-1 and 5-2 are examples of alternative algorithms that have been conceived. They were arrived at from the point of view of control theory.

In the first algorithm, depicted in Figure 5-1, it is assumed that the individual sidelobes of the complex Fourier transform of the object can be modelled by a fairly simple mathematical formula having a small number of free parameters. By curve fitting each lobe of the Fourier modulus (amplitude) to the model, one could determine the parameters and thereby determine the phase. One would first curve fit one larger lobe, compute the magnitude of the model from the fitted parameters, and subtract that model from the modulus measurement. Smaller lobes would be curve-fitted and subtracted resursively from the residual modulus. After all the lobes are modelled, the corresponding phase would be combined with the measured modulus and the image would be computed by inverse Fourier transformation. It is yet to be determined whether the Fourier transform can be modelled as described above.

In the second algorithm, shown in Figure 5-2, the difference in the phase of the Fourier transform of the current estimate and that of the previous estimate is multiplied by a gain factor (K) and added to the previous phase estimate. This is similar to previous iterative



* MODEL CONSISTS OF POLE/ZERO FACTORS OF W_1 TIME POLE/ZERO FACTORS OF W_2

Figure 5-1. Alternative Reconstruction Algorithm - I.

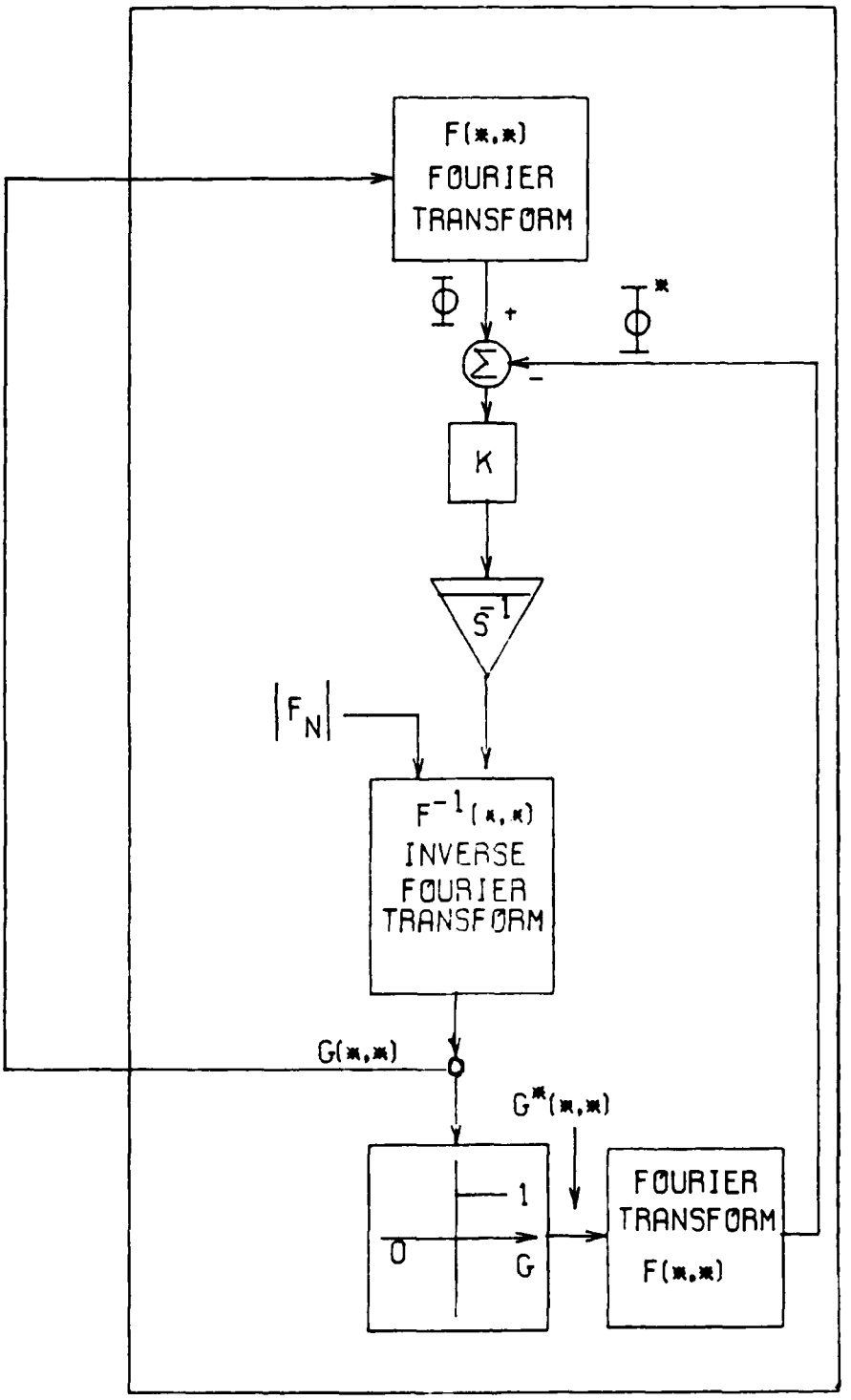


Figure 5-2. Alternative Reconstruction Algorithm - II.

algorithms except that the roles of the two domains are reversed.

Both the methods described above, as well as others, merit further research and implementation.

6
INVESTIGATION OF THE STRIPES PHENOMENON

In a number of cases the iterative Fourier transform algorithm has converged almost all the way to a solution, but then stagnates at an output image that looks like the original object but having a set of stripes superimposed [17, 22] (see Appendix D). In most cases the stripes are of such low contrast as to be hardly noticeable, but occasionally the contrast of the stripes has been high enough to be objectionable. Since the Fourier transform pair is not in perfect agreement with the data and constraints in this condition, the striped images is at a local, rather than the global, minimum of the error, and therefore it does not represent a lack of uniqueness. Although earlier attempts at solving this problem (the stagnation at a striped imaged) failed, we are currently developing methods that will eliminate the stripes.

Since an output image having stripes has a Fourier modulus equal to the measured (assumed to be the correct) Fourier modulus, the stripes must be due to the effects of phase errors. The phase errors must be located in small regions of the Fourier domain in order to produce such a regular striped pattern, with the locations of the regions in the Fourier domain being related to the spatial frequency (spacing) and orientation (angle) of the stripes.

A first attempt at eliminating the stripes was to add noise to the input image after stagnation at a striped output had occurred. The hope was that the added noise would move the solution far enough away from the local minimum so that further iterations of the iterative Fourier transform algorithm would lead to the global minimum rather than falling back into the same local minimum. When this was tried it was found that

after further iterations the algorithm did indeed fall back into the same local minimum, even when the amount of noise added was very large.

A second attempt at solving the stripes problem relied on the knowledge that the phase error was a localized phase error in the Fourier domain. It was found that if a constant phase was added to the phase of the Fourier transform of the object in a given region of the Fourier domain (and in order to preserve the Hermitian property of the Fourier transform, the same constant phase was subtracted in the symmetric region of the Fourier domain), then the corresponding image would look like the original object but with a pattern of stripes superimposed. The resulting synthesized images thus produced had an appearance very much the same as the striped images produced by the iterative algorithm. However, when these images were used as the input to the iterative algorithm, the synthesized stripes immediately went away and the algorithm quickly converged to the true image. This constrasted sharply with the stripes produced by the iterative algorithm, which would not go away. Therefore the phase errors that cause the stripes problem are more complicated than simple constant phase errors over some region of the Fourier domain. This was further shown by attempts to eliminate the stripes by adding various constant phase errors at appropriate regions of the Fourier domain. All such attempts failed to remove the stripes.

A third attempt at removing the stripes involved the addition of Blaschke-like phase functions to the Fourier transform. A Blaschke-like phase is the phase of the unity-modulus function

$$B(u, z) = \frac{1 - u/z^*}{1 - u/z}$$

where u is a Fourier-domain coordinate and z = a + ib. The Blaschke-like phase is global in effect but has its most rapid variation within a region about u = b. Varying phase error corrections of this form were also found to be unsuccessful in solving the stripes problem.

Most recently we have developed two procedures that should solve the stripes problem in most cases. They are based upon two facts:

1. The phase errors that produce the stripes are located in small regions of the Fourier domain, and
2. Output images arrived at by the iterative Fourier transform algorithm started with different initial inputs of random numbers are unlikely to have the same pattern of stripes.

In the first method, three output images are produced by the iterative Fourier transform algorithm using three different initial inputs of random numbers. The three output images are translated so as to be centered at the same point in order to remove any linear phase difference in their Fourier transforms. Then at each point in the Fourier domain, the complex values of the three Fourier transforms are compared. The value whose distance from the other two values is the greatest is discarded and a new value is formed by taking the average of the remaining two (closest) values. In this manner, if in a given region of the Fourier domain one of the three has a phase error (related to the stripes or otherwise), that phase error is eliminated.

In the second method only two different output images need to be produced by the iterative Fourier transform algorithm. Although the stripes are typically of highest contrast where the object is brightest, they also exist outside the known support of the object. So by Fourier transforming the region of the output image having only stripes (outside

the support of the object), the regions of the Fourier domain having the phase error can be identified. Then a new phase estimate is made by using the phase of the Fourier transform of the first output images where it is not influenced by the phase error, and using the phase of the Fourier transform of the second output image where the first was influenced by the phase error.

In both methods above, after the new estimate is formed, further iterations should be performed to allow it to converge closer to a solution.

These methods of correcting the stripes are automatic in the sense that no human judgement or decisions are required during their operation.

Both methods were exercised on a single example and were found to perform very well. Further experimentation with these methods is required to determine their effectiveness in a wider variety of circumstances.

7
EXPERIMENTAL RESULTS USING SHIFT-AND-ADD

Shift-and-add [23, 24] is a method of reconstructing images of astronomical objects from multiple short-exposure images. It consists of shifting all the images so that their maximum values all lie at the same coordinate, then adding (or averaging) them all. This has been shown to work well for objects consisting of a collection of delta functions (points) [23, 24], but it was not demonstrated for realistic extended objects, such as satellites. We implemented the shift-and-add method on the computer and exercised it both on an object consisting of a collection of delta functions and on an extended object.

Figure 7-1(a) shows an object consisting of three delta-function-like points having relative brightness of 100:20:10. Simulated point-spread functions from a telescope including atmospheric turbulence [22] were convolved with the object to arrive at simulated blurred images, an example of which is shown in Figure 7-1(b). Figure 7-1(c) shows the result of shifting and adding 156 blurred images. The three points can be clearly seen, although they are slightly blurred and each is surrounded by a fog. We took shift-and-add one step further by combining it with a form of subtractive deconvolution related to the CLEAN method. An estimate of the effective point-spread function was found by applying shift-and-add to a single point. The outputs from shift and add were then CLEANed by subtracting from the image a version of this point-spread function shifted and scaled to match the peak of the brightest point in the image, and the brightness and position of the peak was noted. The second and third points were CLEANed in a similar fashion in succession. Figure 7-1(d) shows the brightness and positions of the three CLEANed peaks. The brightness and positions are exactly the same (to within an overall translation) as the brightness and

Figure 7-1. Shift-and-Add Algorithm for a Point Object. (a) Original object; (b) image blurred by atmospheric turbulence; (c) output image from shift-and-add; (d) CLEANed version of output image.

positions of the points in the original object. Although no noise was present in the data (but only 156 blurred images were used), this is a very impressive result, demonstrating the power of the shift-and-add method for this type of object.

Figure 7-2 shows a similar experiment for an extended object. The object, shown in Figure 7-2(a), was purposely chosen to be one that satisfies the requirement that it have bright delta-function-like components. That is, it should be one of the easier extended objects for shift-and-add to reconstruct. Two examples of the blurred images of the object are shown in Figures 7-2(b) and (c). The result of using shift-and-add on 156 blurred images is shown in Figure 7-2(d). In this case there is very little information about the original object in the shift-and-add image. Further processing using the CLEAN method did not improve the result.

This one set of experiments was not sufficient to fully delineate the types of objects for which shift-and-add is effective, but we did demonstrate that shift-and-add works very well for an object consisting of a small number of delta-function-like points dominated by a brightest one, but works very poorly for an extended object, even one containing isolated bright points.

Figure 7-2. Shift-and-Add Algorithm for an Extended Object. (a) Original object; (b), (c) images blurred by atmospheric turbulence; (d) output image from shift-and-add.

8
RECONSTRUCTION OF COMPLEX-VALUED OBJECTS USING SUPPORT CONSTRAINT

As discussed elsewhere in this report, for the astronomy problem one has in the object domain a nonnegativity constraint and a weak (loose) support constraint. There are however some problems for which the object is complex-valued or bipolar, precluding a nonnegativity constraint, but for which the support constraint is much stronger (tighter). The same iterative reconstruction algorithm as for the astronomy problem can be used, only without applying the nonnegativity constraint. From the theory of Bruck and Sodin [27], one might expect the solution to usually be unique. For supports known to have certain shapes, such as a triangular shape, the solution is known to be unique (see Appendix B).

Reconstruction experiments using only a support constraint were performed on objects having various support constraints to test the importance of different types of support constraints. Figure 8-1(a) shows an object having triangular support and nonzero values in its three corners. These conditions ensure that the object is uniquely related to its Fourier modulus (see Appendix B). The image shown in Figure 8-1(b) was reconstructed from the Fourier modulus and the a priori knowledge of the triangular support using the iterative Fourier transform algorithm. Nonnegativity was not used although the object happens to be nonnegative. In this case the support was known very precisely. The algorithm converged very rapidly to the correct solution.

Since the uniqueness proof requires the three corners to be nonzero, we wanted to determine the importance of nonzero corners to the iterative algorithm. The same experiment was performed for the object shown in Figure 8-1(c), which is identical to the object shown in Figure

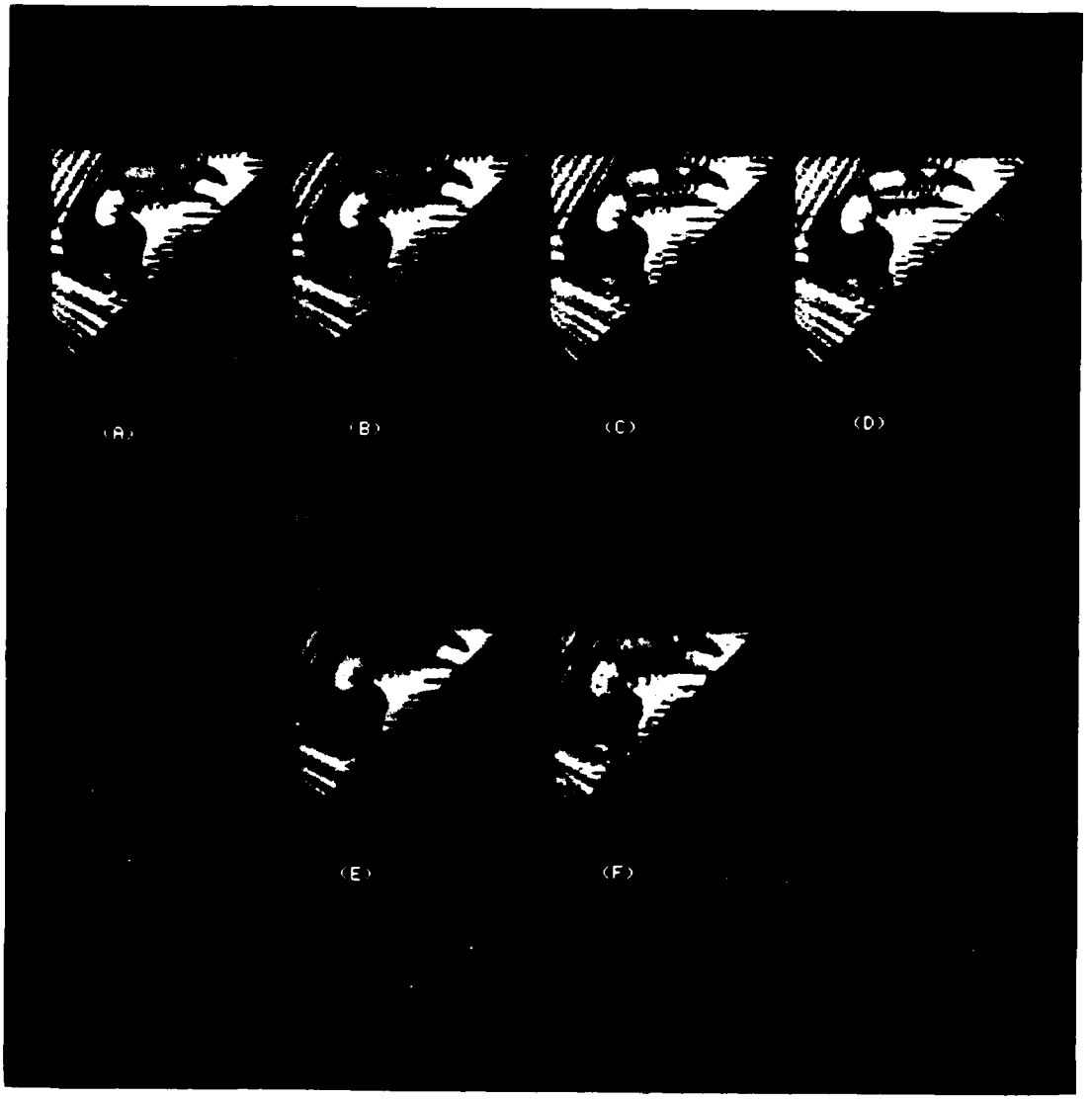


Figure 8-1. Examples of Reconstruction from Fourier Modulus. (a) Object in triangle with bright corners, (b) reconstructed image; (c) object in triangle with zeroed corners, (d) reconstructed image; (e) object in triangle with tapered edges, (f) reconstructed image.

8-1(a) but with the corners zeroed out. The image shown in Figure 8-1(d) was reconstructed from its Fourier modulus and is the correct solution, but more iterations were required in this case than for the object having three bright corners. Therefore the brightness of the corners does play a role, but not a crucial one. The effect of the sharpness of the edges of the object was also investigated. A third object, shown in Figure 8-1(e), which is identical to the object shown in Figure 8-1(c) except that its edges are tapered rather than being abrupt, was formed. The image resulting after over a hundred iterations of the iterative Fourier transform algorithm is shown in Figure 8-1(f). Although the image is very recognizble, it has a noisy appearance. The iterative algorithm found it much more difficult to reconstruct this image than the ones with abrupt or sharp edges. Therefore it appears that edges (although not absolutely essential) are very important to the ability of the iterative algorithm to reconstruct images using only a support constraint in the object domain.

Under an internally funded ERIM program, these results were extended to the case of complex-valued objects for applications such as synthetic-aperture radar (SAR) [25]. Since those results are pertinent here, they will be briefly reviewed. The idea is to have a SAR sensor that does not require an accurate local oscillator, phase-coherent chirp pulses, or compensation of sensor platform motion. This could be done if one could reconstruct the image without accurate knowledge of the phase of the SAR signal history. This might be possible by using the iterative Fourier transform algorithm if a strong support constraint were present. One might have, for example, the ability to illuminate the target area with an illuminaton pattern of known shape or have the far-field pattern of the receive antenna accept reflected radiation from an area of known shape. The modulus information together with the support (beam shape) constraint could then be combined by the iterative Fourier transform algorithm to arrive at a diffraction-limited image.

Figure 8-2 shows an example of a reconstruction experiment [25] of this type. Figure 8-2(a) shows the magnitude of a 64 x 64 pixel sub-area of a complex-valued SEASAT SAR image of an area of land. A binary mask was formed to define the illumination pattern of a hypothetical antenna. The illumination pattern consists of a pair of ellipses, each of 3:1 aspect ratio. A pattern consisting of two separated parts was chosen because theory indicates that the solution is more likely to be unique in that case [13]. The object, the magnitude of which is shown in Figure 8-2(b), was obtained by taking the product of the image shown in Figure 8-2(a) and the illumination pattern. The modulus of the Fourier transform of the object was computed and is shown in Figure 8-2(c). This is equivalent to the modulus of the SAR signal history that would have been collected had the terrain been illuminated by the fixed illumination pattern consisting of the two ellipses. The Fourier modulus was then used together with the known illumination pattern as a support constraint to reconstruct the image which is shown in Figure 8-2(d). The reconstruction was essentially perfect.

Figure 8-3 shows further examples of similar reconstruction experiments performed under the current effort. The goal of this set of experiments was to explore the effects of employing different types of support constraints. Specifically, we wanted to explore what effect the separation of the parts of the support had on the success of the iterative reconstruction algorithm. As shown in Figure 8-2, for widely separated support parts, the iterative algorithm performed very well. Figure 8-3(a) and 8-3(b) show the same object and reconstructed image, respectively, but at a different scale. Figures 8-3(c) and 8-3(d) show a second object and reconstructed image, respectively, for a case in which the separation between the two parts of the support is much smaller. Again the reconstructed image is very faithful. Figure 8-3(e)

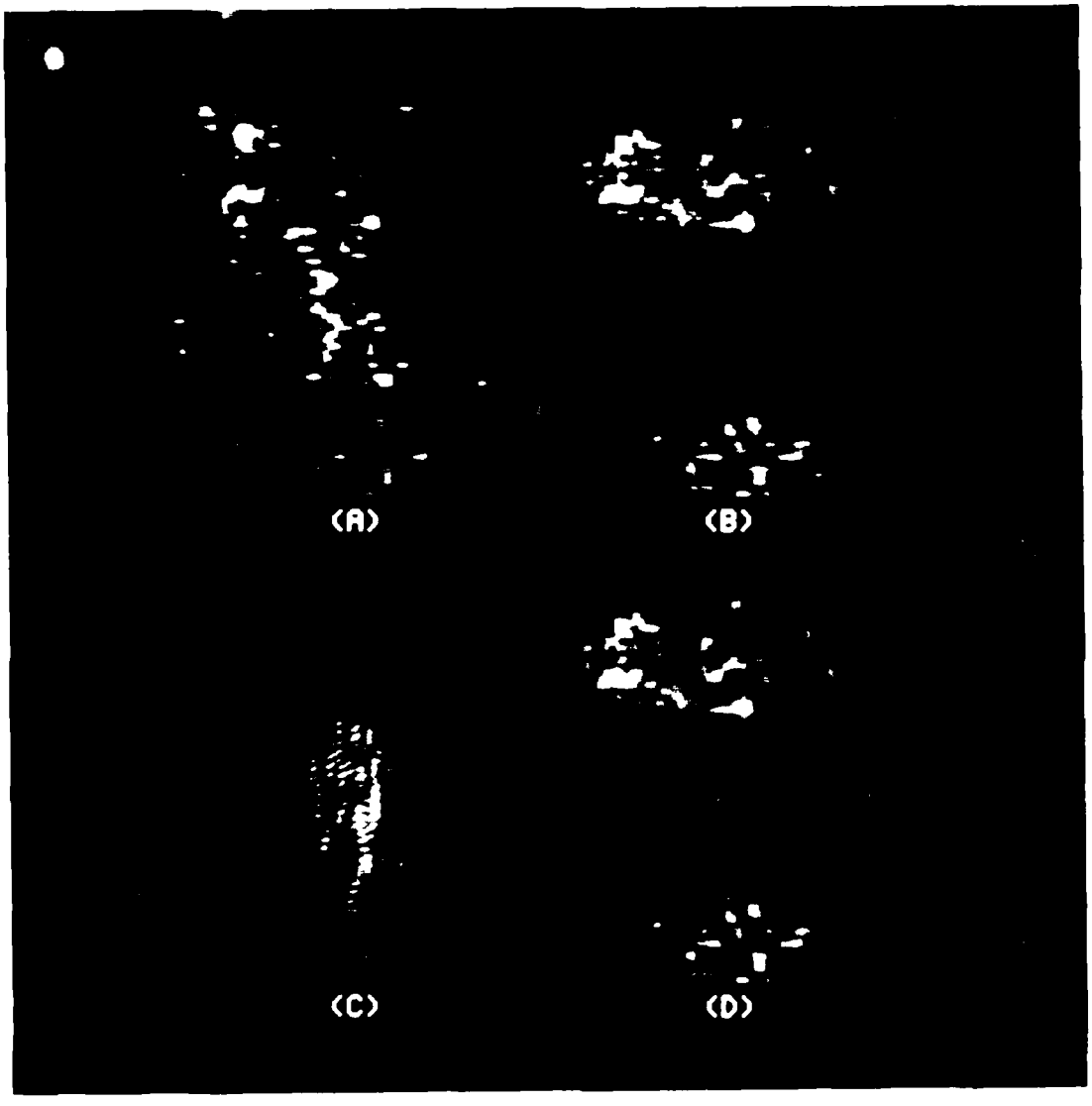


Figure 8-2. Example of Reconstructing a Complex-Valued SAR Image from the Modulus of Its Fourier Transform Using an Illumination-Pattern Support Constraint (a Pair of Ellipses). (a) Magnitude of terrain image with broad illumination pattern; (b) magnitude of ideal terrain image with special illumination pattern; (c) Fourier (phase history) modulus; (d) magnitude of reconstructed image. (Taken from [25].)

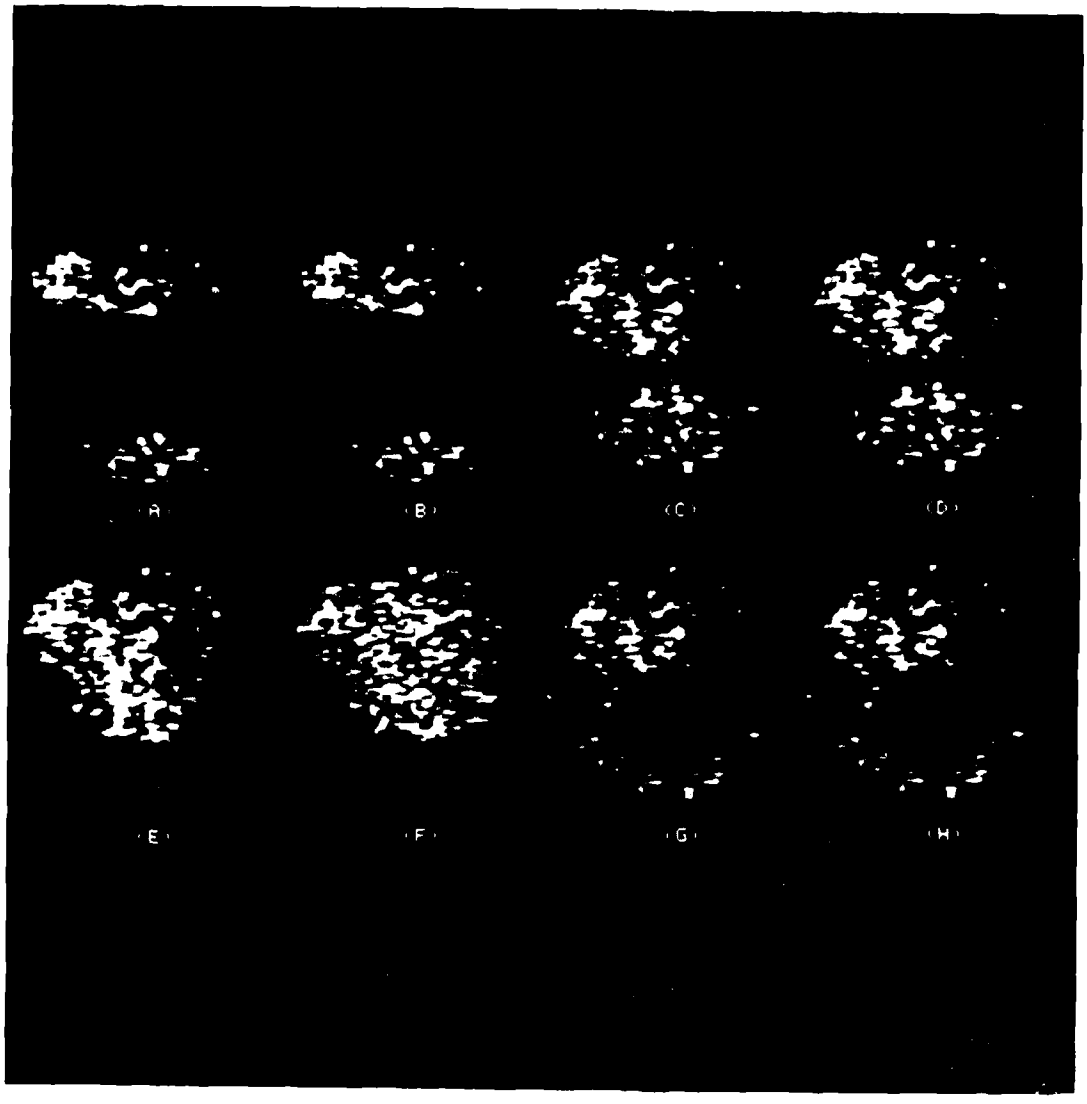


Figure 8-3. Examples of Reconstructing a SAR Image from Modulus of Its Fourier Transform Using Various Support Constraints with the Iterative Transform Algorithm. Illuminated objects: (a), (c), (e), (g); respective reconstructed images: (b), (d), (f), (h).

and 8-3(f) shows a third object and reconstructed image, respectively, for a case in which the two parts overlap (that is, the support is contiguous). In this case the reconstructed image, even after several hundred iterations, does not closely resemble the object, although upon close inspection one can find some features in common. For a fourth case (not shown) in which the support consisted of a single ellipse, the iterative reconstruction algorithm did not produce a recognizable image after several hundred iterations. Figure 8-3(g) and 8-3(h) show a fifth object and reconstructed image, respectively, for a case in which the support was shaped like a donut having a hole offset from the center. In this case the reconstructed image is very faithful. This last example is curious since it does not have separated parts, yet a one-dimensional cut through the center of the support does have separated parts.

The examples of Figure 8-3 demonstrate that the separated nature of the support does have an important effect on the succes of the iterative reconstruction algorithm, and that further investigations along these lines is warranted.

The ability demonstrated above to reconstruct a complex-valued image from the modulus of its Fourier transform using only a support constraint may have very important implicatons for fine-resolution coherent imaging systems such as SAR.

9
REFERENCES

1. A.A. Michelson and F.G. Pease, "Measurement of the Diameter of Alpha Orionis with the Interferometer," Astrophys. J. 53, 249-259 (1921).

2. R. Hanbury Brown and R.Q. Twiss, "Correlation Between Photons in Two Coherent Beams of Light," Nature 177, 27-29 (1956).

3. D.G. Currie, S.L. Knapp, and K.M. Liewer, "Four Stellar-Diameter Measurements by a New Technique: Amplitude Interferometry," Astrophys. J. 187, 131-144 (1974).

4. A. Labeyrie, "Attainment of Diffraction Limited Resolution in Large Telescopes by Fourier Analysing Speckle Patterns in Star Images," Astron. and Astrophys. 6, 85-87 (1970).

5. D.Y. Gezari, A. Labeyrie, and R.V. Stachnik, "Speckle Interferometry: Diffraction-Limited Measurements of Nine Stars with the 200-Inch Telescope," Astrophys. J. Lett. 173, L1-L5 (1972).

6. J.R. Fienup, "Reconstruction of an Object from the Modulus of Its Fourier Transform," Opt. Lett. 3, 27-29 (1978).

7. J.R. Fienup, "Space Object Imaging through the Turbulent Atmosphere," Opt. Eng. 18, 529-534 (1979).

8. J.R. Fienup, "Phase Retrieval Algorithms: A Comparison," Appl. Opt. 21, 2758-2769 (1982).

9. J.R. Fienup, "Reconstruction and Synthesis Applications of an Iterative Algorithm," in Transformations in Optical Signal Processing, W.T. Rhodes, J.R. Fienup and B.E.A. Saleh, eds., Proc. SPIE 373, 147-160 (1981).

10. J.R. Fienup, "Reconstruction of Objects Having Latent Reference Points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

11. J.R. Fienup, "Diffraction-Limited Imaging of Space Objects I," Interim Scientific Report to AFOSR, ERIM Rept. No. 161900-6-T (May 1983).

12. J.R. Fienup, "Ambiguity of Phase Retrieval Using Boundary Conditions," submitted to J. Opt. Soc. Am.

13. T.R. Crimmins and J.R. Fienup, "Uniqueness of Phase Retrieval for Functions with Sufficiently Disconnected Support," J. Opt. Soc. Am. 73, 218-221 (1983).

14. J.R. Fienup, "Phase Retrieval in Astronomy," in Signal Recovery and Synthesis with Incomplete Information and Partial Constraints, digest of papers (Opt. Soc. Am., 1983), Th A8, Jan. 12-14, 1983, Incline Village, NV.

15. J.R. Fienup, "Two-Dimensional Image Reconstruction Algorithms," in Information Processing in Astronomy and Optics, digest of papers (Opt. Soc. Am., 1983), Th A11, June 23-24, 1983, St. Paul, MN.

16. J.R. Fienup, "Holographic Reconstruction with Latent Reference Points," presented at the Annual Meeting of the Optical Society of America, New Orleans, LA, October 1983; Abstract: J. Opt. Soc. Am. 73, 1861 (1983).

17. J.R. Fienup, "Experimental Evidence of the Uniqueness of Phase Retrieval from Intensity Data," in Indirect Imaging, Proc. URSI/IAU Symposium, 30 Aug.-1 Sept., 1983, Sydney, Australia (Cambridge University Press, 1984), pp. 99-109.

18. J.R. Fienup, "Comments on 'The Reconstruction of a Sequence from the Phase or Magnitude of Its Fourier Transformation'," IEEE Trans. Acoustics, Speech, Signal Processing ASSP-31, 738-739 (1983).

19. J.R. Fienup, "Fine Resolution Imaging of Space Objects," Final Scientific Report to AFOSR, Contract No. F49620-80-C-0006, ERIM Report No. 145400-14-F, February 1982.

20. M.A. Fiddy, B.J. Brames, and J.C. Dainty, "Enforcing Irreducibility for Phase Retrieval in Two Dimensions," Opt. Lett 8, 96-98 (1983).

21. R.H.T. Bates and W.R. Fright, "Composite Two-Dimensional Phase-Restoration Procedure," J. Opt. Soc. Am. 73, 358-365 (1983).

22. G.B. Feldkamp and J.R. Fienup, "Noise Properties of Images Reconstructed from Fourier Modulus," in 1980 International Optical Computing Conference, Proc. SPIE 231, 84-93 (1980).

23. R.H.T. Bates and F.M. Cady, "Towards True Imaging by Wideband Speckle Interferometry," Opt. Commun. 32, 365-369 (1980).

24. F.M. Cady and R.H.T. Bates, "Speckle Processing Gives Diffraction-Limited True Images from Severely Aberrated Instruments," Opt. Lett. 5, 438-440 (1980).

25. J.R. Fienup, "Intensity Only Sensor," ERIM Report No. 659105-1-X (November 1983).

26. M.H. Hayes and T.F. Quatieri, "Recursive Phase Retrieval Using Boundary Conditions," J. Opt. Soc. Am. 73, 1427-1433 (1983).

27. Yu. M. Bruck and L.G. Sodin, "On the Ambiguity of the Image Reconstruction Problem," Opt. Commun. 30, 304-308 (1979).

Appendix A

RECONSTRUCTION AND SYNTHESIS APPLICATIONS OF AN ITERATIVE ALGORITHM

J.R. Fienup

Reprinted from W.T. Rhodes, J.R. Fienup and B.E.A. Saleh, Eds., Transformations in Optical Signal Processing, Proc. SPIE 373, 147-160 (1981) (published in 1984).

Reconstruction and synthesis applications of an iterative algorithm

J. R. Fienup
Environmental Research Institute of Michigan
P.O. Box 8618
Ann Arbor, Michigan 48107

Abstract. This paper reviews the Gerchberg-Saxton algorithm and variations thereof that have been used to solve a number of difficult reconstruction and synthesis problems in optics and related fields. It can be used on any problem in which only partial information (including both measurements and constraints) of the wavefront or signal is available in one domain and other partial information is available in another domain (usually the Fourier domain). The algorithm combines the information in both domains to arrive at the complete description of the wavefront or signal. Various applications are reviewed, including synthesis of Fourier transform pairs having desirable properties as well as reconstruction problems. Variations of the algorithm and the convergence properties of the algorithm are discussed.

1. INTRODUCTION

There exist many problems that are very difficult to solve in astronomy, x-ray crystallography, electron microscopy, spectroscopy, wavefront sensing, holography, particle scattering, superresolution, radar signal and antenna synthesis, filter design, and other disciplines that share an important feature. These are problems that involve the reconstruction or synthesis of a wavefront (or an object or a signal, etc.) when partial information or constraints exists in each of two different domains. The second domain is usually the Fourier transform domain. This paper describes a method of combining all the available information in the two domains to arrive at a complete description, thereby solving the problems.

The problems fall into two general categories: (1) reconstruct the entire information about a function (an image, wavefront, signal, etc.) when only partial information is available in each of two domains; and (2) synthesize a (Fourier) transform pair having desirable properties in both domains. A reconstruction problem arises when only partial information is measured in one domain, and in the other domain either partial information is measured or certain constraints are known *a priori*. The information available in any one domain is insufficient to reconstruct the function or its transform. A synthesis problem typically arises when one wants the transform of a function to have certain desirable properties (such as uniform spectrum, low sidelobes, etc.) while the function itself must satisfy certain constraints or have certain desirable properties. Because arbitrary sets of properties and constraints can be contradictory, there may not exist a transform pair that is completely desirable and satisfies all the constraints. Nevertheless, one seeks a transform pair that comes as close as possible to having the desirable properties and satisfying the constraints in both domains.

Both the reconstruction and the synthesis problems can be expressed as follows, if the meaning of the word "constraints" is broadened to include any kind of measured data, desirable properties, or *a priori* conditions:

> *Given a set of constraints placed on a function and another* set of constraints placed on its transform, find a transform pair (i.e., a function and its transform) that satisfies both sets of constraints.

Once a solution is found to such a problem, the question often remains: is the solution unique? For synthesis problems, the uniqueness is usually unimportant—one is satisfied with any solution that satisfies all the constraints; often a more important problem is whether there exists *any* solution that satisfies what may be arbitrary and conflicting constraints. For reconstruction problems, the uniqueness properties of the solution are of central importance. If many different functions satisfying the constraints could give rise to the same measured data, then a solution that is found could not be guaranteed to be the correct solution. The question of uniqueness must be studied for each problem. Fortunately, as will be described later, for some important reconstruction problems the solution usually is unique.

An effective approach to solving the large class of problems described above is the use of iterative algorithms related to the Gerchberg-Saxton algorithm.[1] The algorithms involve the iterative transformation back and forth between the two domains, with the known constraints applied repetitively in each domain.

The basic algorithm is presented in Sec. 2. A number of different applications having different types of constraints are described, and examples are shown in Sec. 3. In Sec. 4 the convergence properties of the algorithm are discussed, and improved versions of the algorithm are reviewed. A brief summary and comments are included in Sec. 5.

2. THE BASIC ITERATIVE ALGORITHM

The first published account of the iterative algorithm was its use by Gerchberg and Saxton[1] to solve the electron microscopy problem. For this problem both the modulus (magnitude) of a complex-

valued image and the modulus of its Fourier transform are measured, and the goal is to reconstruct the phase in both domains. Apparently unknown to Gerchberg and Saxton, the method was invented somewhat earlier by Hirsch, Jordan, and Lesem[2] to solve a synthesis problem for computer-generated holograms that has a similar set of constraints. (This will be described later in more detail.) The method was again reinvented for a similar problem in computer holography by Gallagher and Liu.[3] The fact that the algorithm was invented repeatedly testifies to its simplicity and effectiveness.

2.1. Gerchberg-Saxton algorithm

In what immediately follows, the iterative algorithm is described in terms of its application to the electron microscopy reconstruction problem. An excellent treatment of the electron microscopy phase problem and its solution by this and other methods can be found in Ref. 4. Later it is shown how to apply the same principles to a large class of problems.

Suppose that the electron wave function in an image plane is described by the two-dimensional (2-D) complex-valued function

$$f(x) = |f(x)|\, e^{i\psi(x)}. \tag{1}$$

Its Fourier transform, the wave function in a far-field diffraction plane, is given by

$$F(u) = |F(u)|\, e^{i\theta(u)} = \mathcal{F}[f(x)] = \int_{-\infty}^{\infty} f(x) e^{-i2\pi u \cdot x}\, dx, \tag{2}$$

where x and u are the vector coordinates in the spatial (image) domain and the spatial frequency (far-field diffraction) domain, respectively. The notation used throughout this paper is that functions represented by capital letters are the Fourier transforms of the functions represented by the corresponding lower-case letters. It is assumed that the intensity spatial distributions are measured in each domain, but the phase information is lost. Therefore, one wishes to reconstruct $\psi(x)$ and $\theta(x)$ from $|f(x)|$ and $|F(u)|$.

The iterative algorithm for solving this problem is depicted in Fig. 1. One iteration (the k^{th} iteration) of the algorithm proceeds as follows. A trial solution for the wave function (an estimate of the wave function), $g_k(x)$, is Fourier transformed yielding

$$G_k(u) = |G_k(u)| \exp[i\phi_k(u)] = \mathcal{F}[g_k(x)]. \tag{3}$$

Then a new Fourier-domain function, $G_k'(u)$, is formed by replacing the computed Fourier modulus by the measured Fourier modulus, $|F(u)|$, and keeping the computed phase:

$$G_k'(u) = |F(u)| \exp[i\phi_k(u)]. \tag{4}$$

The resulting $G_k'(u)$, which is in agreement with all the known measurements and constraints in the Fourier domain, is inverse Fourier transformed, yielding the wave function $g_k'(x)$. The iteration is completed by forming a new estimate for the wave function, $g_{k+1}(x)$, which is obtained by replacing the computed modulus of $g_k'(x)$ with the measured modulus $|f(x)|$, and keeping the computed phase.

The algorithm consists of no more than enforcing what information is available on the wave function, Fourier transforming, imposing what information is available on the wave function's Fourier transform, inverse transforming, and repeating these simple operations for a number of iterations. What makes the algorithm practical is the existence of a fast Fourier transform[5] (FFT), so that the number of computations per iteration goes only as NlogN, where N is the number of samples of the function computed. This compares very favorably with some other iterative methods, such as Newton-Raphson,[4] for which the number of computations per iteration goes as N^3.

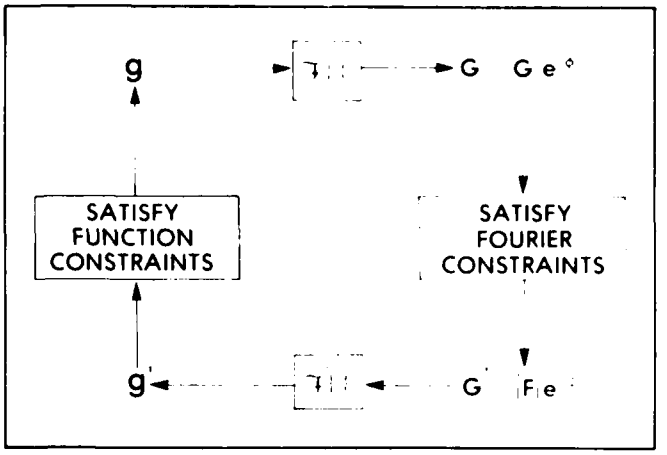


Fig. 1. Block diagram of the iterative error-reduction algorithm.

A measure of the progress of the iterations, and a criterion by which one can determine when a solution has been found, is the normalized mean-squared error, which is defined in the Fourier domain by

$$E_F^2 = \frac{\int_{-\infty}^{\infty} [\,|G_k(u)| - |F(u)|\,]^2\, du}{\int_{-\infty}^{\infty} |F(u)|^2\, du} \tag{5}$$

or in the image domain by

$$E_0^2 = \frac{\int_{-\infty}^{\infty} [\,g_k'(x) - f(x)\,]^2\, dx}{\int_{-\infty}^{\infty} |f(x)|^2\, dx}. \tag{6}$$

It has been shown that the algorithm converges in the sense that the mean-squared error can only decrease at each iteration.[1,4,6] The issue of convergence will be discussed in greater detail in Sec. 4.

2.2. Error-reduction iterative algorithm

It is now known that with slight modifications this same algorithm can be applied to many different problems having a variety of available constraints or measurements.[7] Let the function f(x) represent a wavefront, an object, a signal, an antenna array, a spectral density function, an electron density function, etc., where x is an N-dimensional vector (spatial, angular, time, etc.) coordinate. Depending on the problem, f(x) may be complex valued or real valued and, if real, may or may not be nonnegative. Its Fourier transform, F(u), is given by Eq. (2) and is complex valued for most problems. The N-dimensional vector u is a (spatial, angular, time, etc.) frequency coordinate. One can instead consider another transformation of f(x), such as the Fresnel transform, which has been used for more than one problem.[2,8,9] For simplicity of discussion, the Fourier transform will be assumed, but the reader should keep in mind that what is said also applies to a number of other transformations as well (although the method becomes less attrac-

tive if a fast transform algorithm is not available).

With only slight modifications, the Gerchberg-Saxton algorithm can be used to solve the wide class of problems described in Sec. 1. *Referring again to the block diagram of the algorithm* in Fig. 1, all that is required is to impose constraints in each domain that are pertinent to the problem of interest. At the k^{th} iteration, $g_k(x)$, an estimate of f(x), is Fourier transformed, yielding $G_k(u)$, which is given by Eq. (3). Then a new Fourier-domain function $G_k'(u)$ is formed from $G_k(u)$ by making the smallest possible changes in $G_k(u)$ that allow it to satisfy the Fourier-domain constraints. For example, if the Fourier-domain constraint is that the Fourier modulus equals $|F(u)|$ over some region of the Fourier domain, then $|F(u)|$ is substituted for $|G_k(u)|$ in that region. The new Fourier-domain function $G_k'(u)$, which satisfies the Fourier-domain *constraints*, is inverse Fourier transformed to yield $g_k'(x)$. To complete one iteration, a new estimate $g_{k+1}(x)$ is formed from $g_k'(x)$ by making the smallest possible changes in $g_k'(x)$ that allow it to satisfy the function-domain constraints. One example is that if the function is complex valued and it is constrained to have a modulus equal to $|f(x)|$ over some region of space, then $|f(x)|$ is substituted for $g_k'(x)$ in that region. A special case of this is when the function is to be zero outside a certain interval (the Fourier function is bandlimited). Another example is that if the function is constrained to be nonnegative, then $g_{k+1}(x)$ is set equal to $g_k'(x)$ for those x where $g_k'(x) \geq 0$, and $g_{k+1}(x)$ is set equal to zero for those x where $g_k'(x) < 0$. In summary, one transforms back and forth between the two domains, forcing the function to satisfy the constraints in each domain.

For reconstruction problems, whatever characteristics of the actual F(u) and f(x) that are measured or are known *a priori* are imposed on $G_k(u)$ and $g_k'(x)$, respectively. For synthesis problems, one imposes on $G_k(u)$ and $g_k'(x)$ whatever characteristics one might desire F(u) and f(x), respectively, to have. Once the constraints are defined, the algorithm proceeds the same for synthesis problems as for reconstruction problems. In fact, there are some synthesis problems that are mathematically indistinguishable from some reconstruction problems, and they are handled identically by the algorithm.

The first iteration of the algorithm can be started in a number of ways, for example, by setting $g_1(x)$ or $\phi_1(x)$ equal to an array of random numbers. The iterations continue until a Fourier transform pair is found that satisfies all the constraints in both domains to within the desired accuracy (or, if convergence is too slow, until one loses interest or the money runs out). The mean-squared error can generally be defined in the Fourier domain by

$$E_1^2 = \frac{\int_{-\infty}^{\infty} |G_k(u) - G_k'(u)|^2 \, du}{\int_{-\infty}^{\infty} |G_k'(u)|^2 \, du} \qquad (7)$$

or in the function domain by

$$E_0^2 = \frac{\int_{-\infty}^{\infty} |g_{k+1}(x) - g_k'(x)|^2 \, dx}{\int_{-\infty}^{\infty} |g_k(x)|^2 \, dx} . \qquad (8)$$

In each of these two expressions, the integrand in the numerator is the squared modulus of the amount by which the computed function violates the constraints in that domain. It is easily seen that these expressions reduce to Eqs. (5) and (6), respectively, for the electron microscopy problem.

Just as in the electron microscopy problem, for problems having other sets of constraints it will be shown in Sec. 4 that the algorithm converges, that is, the error decreases at each successive iteration. The algorithm depicted in Fig. 1 may be referred to as the "error-reduction" algorithm for that reason, as well as to distinguish it from algorithms described in Sec. 4 that are related to it but converge faster. Typically, the error is reduced very rapidly for the first few iterations of the error-reduction algorithm, but more slowly for later iterations. For some applications, the error-reduction algorithm has been very successful in finding solutions using a reasonable number of iterations. However, for some other applications, the mean-squared error decreases extremely slowly with each iteration, and an impractically large number of iterations is required. The improved algorithms described in Sec. 4 do much to alleviate this problem.

2.3. Alternative descriptions of the algorithm

Once a solution (i.e., a Fourier transform pair satisfying all the constraints in both domains) is found, the error-reduction algorithm ceases to make changes to the estimate, and the algorithm locks on to the solution. The operations of enforcing the constraints in each domain would then leave the function estimate and its Fourier transform unaltered, since they already satisfy the *constraints*. Now let us define the operation S[g(x)] as the successive Fourier transformation of g(x), followed by the imposition of the Fourier domain constraints, followed by inverse Fourier transformation, followed by imposition of the object domain constraints. That is, the operation S is just the performance of one iteration of the error-reduction algorithm, and

$$g_{k+1}(x) = S[g_k(x)] . \qquad (9)$$

From the discussion above, it is evident that any solution f(x) must *satisfy the relation*

$$f(x) = S[f(x)] . \qquad (10)$$

When presented in this form, it is seen that the error-reduction algorithm is a particular implementation of the method of successive approximations.[10]

The method of successive approximations can be more easily understood from the following simple example. Suppose one wishes to solve the following equation for y:

$$4y^4 - 4y + 1 = 0 . \qquad (11)$$

Based on the relation $y = y^4 + 1/4$, one could write

$$y_{k+1} = S_1(y_k) = y_k^4 + 1/4 . \qquad (12)$$

Using the method of successive approximations to find the solution, one would pick an initial estimate, say $y_0 = 0.1$, and employing Eq. (12) compute $y_1 = 0.2501$, $y_2 = 0.2539$, etc., and rapidly converge to the solution $y' = 0.2541737\ldots$. However, it converges to y' only for $y_0 < y'' = 0.8967902\ldots$. For $y_0 > y''$, Eq. (12) diverges; and for $y_0 = y''$, it stays at y'', the second solution. On the other hand, one could just as logically have chosen

$$y_{k+1} = S_2(y_k) = (y_k - 1/4)^{1/4} . \qquad (13)$$

This second form converges to the second solution y'' for $y_0 > y''$, diverges for $y_0 < y''$, and stays at y' for $y_0 = y'$. Figure 2, a graphical representation of Eq. (12), shows the two solutions, y' and y''. The irregular staircase between the two curves y and $y^4 + 1/4$

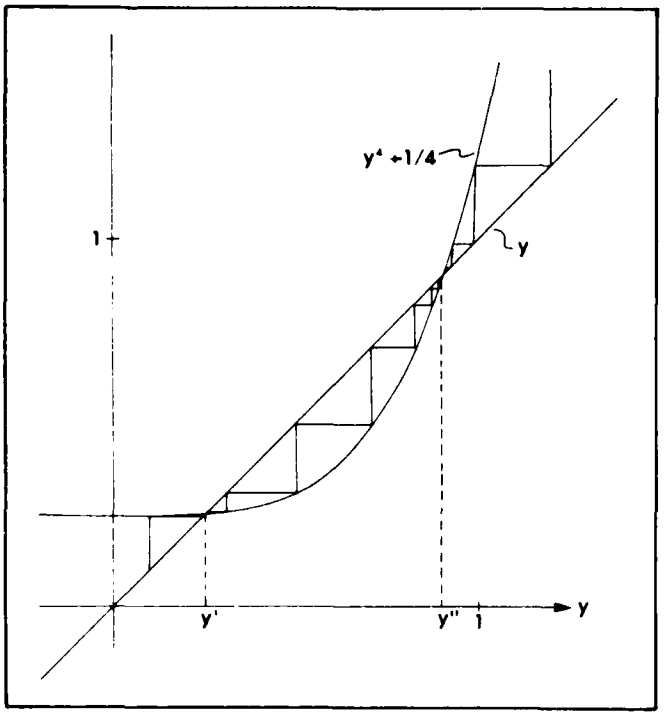


Fig. 2. Method of successive approximations for solving $4y^4 - 4y + 1 = 0$.

1/4 indicates how the estimate y_k approaches the two solutions. Criteria on the derivative of S(y) determine whether the algorithm converges.[11]

The error-reduction algorithm, as described by Eqs. (9) and (10), is analogous to the example of successive approximations described above, except that instead of operating on a scalar y, it operates on a function g(x). As seen from the example, the method of successive approximations may or may not converge, depending on the particular form chosen and on the initial estimate. Fortunately, as will be discussed further in Sec. 4, the error-reduction algorithm never diverges. It may, however, stagnate. A simple example of stagnation of the method of successive approximations is shown by the following. In solving $x = 2 - x$ (which has the obvious solution $x = 1$), starting with the initial estimate x_0, one obtains $x_1 = 2 - x_0$, $x_2 = 2 - (2 - x_0) = x_0, \ldots, x_{2k-1} = 2 - x_0, x_{2k} = x_0$, etc., and no progress is made toward the solution.

Another way of understanding the error-reduction algorithm, applicable for certain sets of constraints, is the alternating projection of the function onto specified subspaces in a Hilbert space.[12] This, along with the possibility of closed-form solutions,[13] is discussed in the contribution to this volume by Marks and Smith.

3. APPLICATIONS

A large number of important problems in optics and related fields fit the problem description in Sec. 1 and can be solved by the iterative algorithm (by the error-reduction algorithm described in Sec. 2 and the related algorithms described in Sec. 4). One particular application, that of spectral extrapolation or superresolution, is discussed in detail in the contribution to this volume by Marks and Smith. In this section, several classes of applications are listed, followed by more detailed discussions of some of the applications, including examples.

In Sec. 1, a distinction was made between reconstruction problems and synthesis problems. Another useful way to classify such problems is according to the type of information available. For one set of problems, the modulus (magnitude or amplitude) of a complex-valued function and the modulus of its Fourier transform are measured (or are given), and one wishes to know the phase of the Fourier transform pair in both domains. These include the phase retrieval problem in electron microscopy, the phase retrieval problem in wavefront sensing, the design optimization of radar signals and antenna arrays having desirable properties, and phase coding and spectrum shaping problems for computer-generated holograms and other applications. These applications often involve the Fresnel transform for the near-field case instead of the Fourier transform.

For another set of problems, the function is known to be real and nonnegative and the modulus of its Fourier transform is measured. These include the phase problems of x-ray crystallography, Fourier transform spectroscopy, imaging through atmospheric turbulence using interferometer data, and pupil function determination.

For another set of problems, a low-resolution (i.e., a low-pass filtered) version of a function is measured (i.e., its complex Fourier transform is measured only over a certain interval), and the function is known to have a finite extent (i.e., it is zero outside of some known region of support). This is the spectral extrapolation or superresolution problem for band-limited time signals or for imaging of objects of finite extent.

For another set of problems, the function is known to be nonnegative and of finite extent and its complex Fourier transform is measured only over a partially filled aperture. These include the interpolation of the complex visibility function for long baseline radio interferometry and the missing-cone problem in x-ray tomography.

For still another set of problems, the modulus of a complex-valued function is given, and one wishes to find an associated phase function that results in a Fourier transform whose complex values fall on a prescribed set of quantized complex values. These include the reduction of quantization noise in computer-generated holograms and in coded signal transmission.

Another problem is to reconstruct the modulus of a complex-valued function from the phase of the function, given the fact that the Fourier transform of the function has finite support.

The number of types of problems solvable by the iterative algorithm appears to be limited only by one's ingenuity in defining different combinations of information that might be available in each of two domains.

3.1. Modulus—modulus constraints

3.1.1. Electron microscopy

Among the applications for which the modulus is given in each of two domains, the electron microscopy phase retrieval problem was one of the earliest applications of the error-reduction algorithm and has been the problem most heavily investigated.[1,4,8,14,15] The error-reduction (Gerchberg-Saxton) algorithm has been shown to perform very successfully for this problem, and the solution is usually unique.[15] The reader is referred to a book by Saxton[4] for a thorough review.

3.1.2. Spectrum shaping

A second application for which the modulus is given in each of two domains is the spectrum shaping problem. Spectrum shaping is a synthesis problem that can be stated as follows: given the modulus $f(x)$ of a complex-valued wavefront, $g(x) = f(x) \exp[i\theta(x)]$, find a phase function $\theta(x)$ such that $\mathcal{F}[g(x)]$ is equal to a given spectrum $F(u)$. Such a problem is the one suggested by the Escher engraving shown in Fig. 3, in which a bird transforms into a fish. One wishes to find a function with modulus being a picture of a fish, which has a Fourier transform with modulus being a picture of a bird. Or, in terms of computer holography, find a phase function to assign to the image of a fish so that the hologram will look like an image of a bird. Figure 4(a) shows the actual "bird" and "fish" binary patterns used for our experiment. For the first iteration, the fish object was random phase coded, Fourier transformed, and the modulus of the Fourier transform was replaced with the modulus of the bird pattern shown in Fig. 4(a). The result was inverse

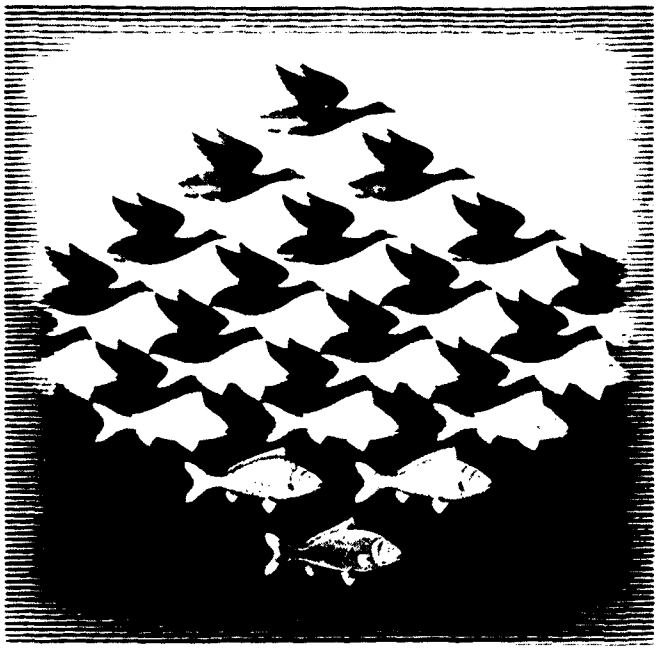


Fig. 3. Bird transforms into fish ("Sky and Water" by M. C. Escher). This reproduction was authorized by the M. C. Escher Foundation, The Hague, Holland/G.W. Breughel.

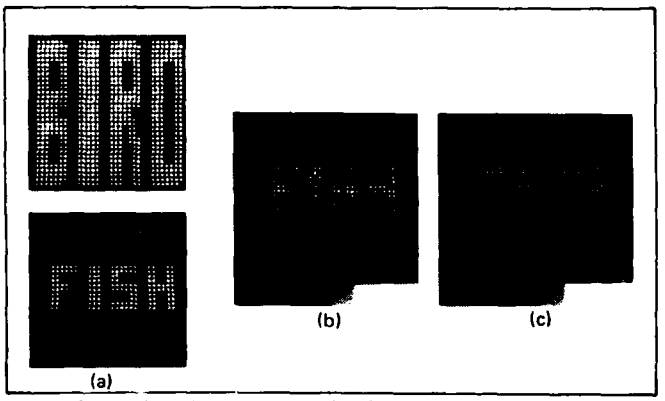


Fig. 4. Example of spectrum shaping. (a) Bird hologram and desired fish image; (b) fish output image after random phase coding of input; (c) output image after seven iterations of the iterative algorithm.

Fourier trsformed, yielding the very noisy output image shown in Fig. 4(b). The iterative algorithm was then used for seven iterations, resulting in the improved image shown in Fig. 4(c). For this example, increasing the number of iterations resulted in a further improvement of the quality of the image; that is, a Fourier transform pair was found that more closely satisfied the constraints in both domains.

Spectrum shaping is also important in computer holography for reducing quantization noise. The objective of computer holography[16] is to synthesize a transparency that can modulate a wavefront according to a calculated wavefront, often corresponding to Fourier coefficients (or samples of the Fourier transform of an image) computed by the discrete Fourier transform. Let $F = \mathcal{F}[f]$ be the desired wavefront modulation and f be the complex-valued function describing the desired image. Due to the limitations of the recording devices and materials used to synthesize computer holograms, it is often not possible to represent exactly any arbitrary complex Fourier coefficient. An extreme example of this is the

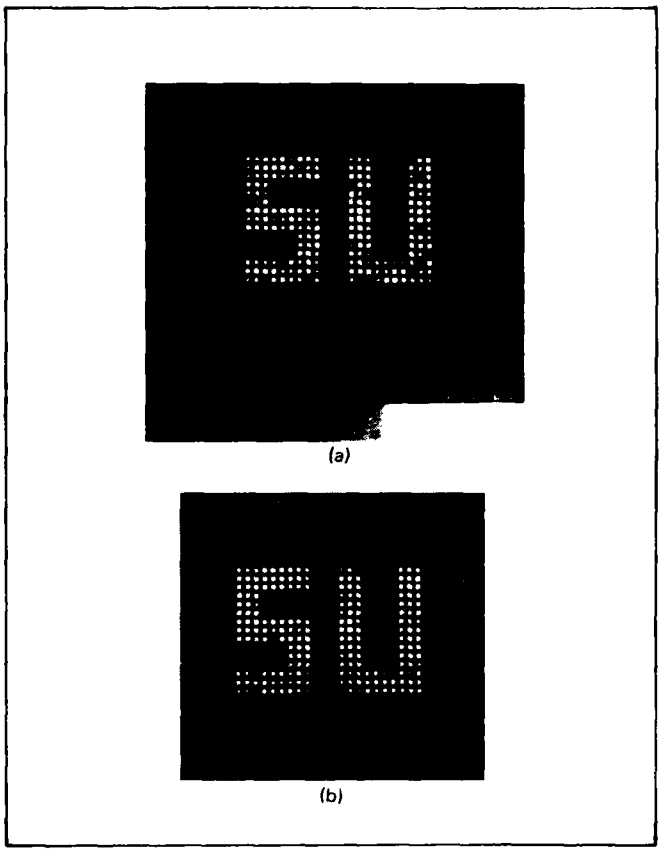


Fig. 5. Computer-simulated images from kinoform. (a) object random phase coded; (b) after eight iterations of the iterative algorithm.

kinoform,[17] which allows nearly continuous phase control by varying the thickness of the recording medium, but which quantizes the modulus to a single level. (If the gray-level recording device used to synthesize a kinoform has a finite number of gray levels, then the phase is quantized as well.) The desired coefficient F is only approximated by the quantized value F/ F . Since only the squared modulus (the intensity) of the image is observed, one is free to choose the phase of the object (phase code the object) in such a way as to reduce the variance (dynamic range) of F . In this way the quantization noise in kinoforms and, to a lesser extent, in other types of computer-generated holograms can be greatly reduced. Random phase and various deterministic phase codes[18] cause considerable reduction in the variance of F , but substantial errors remain.[19]

It was for the kinoform application that the iterative algorithm was first invented.[2,3] Figure 5 shows an example of its use for this synthesis problem.[7] Figure 5(a) shows the image resulting when the input image was random phase coded, encoded as a kinoform in the Fourier plane, and reconstructed by inverse Fourier transformation. The ideal image would be the binary (= 0 or 1) block letters SU. Figure 5(b) shows the improved result after eight iterations of the iterative algorithm. In this case, the image-domain constraint is that the modulus equal the SU pattern, and the Fourier-domain constraint is that the modulus equal a constant.

A problem very similar to the kinoform problem is that of synthesizing a quasi-random radar signal having good autocorrelation properties. Specifically, one would like to synthesize a radar signal f(t) which is a pure phase function, i.e., f(t) = 1, over some interval of time and which has an autocorrelation function which approaches a delta-function, i.e., its Fourier spectrum $|F(\nu)|^2$ is constant over the bandwidth of interest. From the examples shown

above, it is obvious that the iterative method would be an effective tool for synthesizing such radar signals.

Another spectrum-shaping application is the phasing of elements of an array of antennas in order to achieve a far-field pattern having desirable properties. For example, one might wish to phase the antenna elements in such a way as to minimize the maximum sidelobe of the far-field pattern or to place nulls of the antenna pattern at several different prescribed locations simultaneously. A related application for which the iterative method has been used is the transformation of a Gaussian laser beam into a beam having a more nearly rectangular profile.[20]

3.1.3. Wavefront sensing

The wavefront sensing application is very similar to the electron microscopy problem. Suppose that one measures the image $|f(x)|^2$ of a point source using an aberrated optical system, where the aberrations may be due to atmospheric turbulence or due to the optical system itself. Assuming that the aberration is a pure phase function, then F(u), the Fourier transform of f(x), has modulus $|F(u)|$ equal to the aperture function of the optical system. The problem is to reconstruct the phase of F(u) given $|F(u)|$ and $|f(x)|$. Several investigators[9,21,22] have applied the error-reduction algorithm to this problem with generally good results.

3.2. Nonnegativity—modulus constraints

For some reconstruction problems, the physical quantity of interest can be represented as a nonnegative function, and one is able to measure only the modulus of its Fourier transform (or at least the measured modulus information has a much higher signal-to-noise ratio than the measured phase). From the Fourier modulus, one wishes to reconstruct the Fourier phase or, equivalently, the function itself. Since the autocorrelation of the function is available as the inverse Fourier transform of the squared Fourier modulus,[23] this problem is equivalent to reconstructing the function from its autocorrelation. This problem, referred to as the phase retrieval problem of optical coherence theory, arises in spectroscopy,[24] a one-dimensional problem; in astronomy, a two-dimensional problem; and in x-ray crystallography,[25] a three-dimensional problem. In spectroscopy, the nonnegative spectral density, $g(\nu)$, is the Fourier transform of the complex degree of temporal coherence, $\gamma(\tau)$, of which $|\gamma(\tau)|$ is most easily measured. In x-ray crystallography, the nonnegative electron density function, $\varrho(x, y, z)$, which is periodic, is the Fourier transform of the structure factor F_{hkl}, of which $|F_{hkl}|$ is measured by a diffractometer. The astronomy problem will be described in more detail later.

3.2.1. Uniqueness of solutions

For the one-dimensional problem, use of the iterative algorithm (or any other method) to reconstruct the function from its Fourier modulus is of limited interest since the solution in the general case is usually not unique.[26,27] The uniqueness of the solution for the one-dimensional problem can be analyzed using the theory of analytic functions, from which one finds that additional solutions can be generated by "flipping zeros" of the Fourier transform analytically extended over the complex plane.[26,27] The additional "solutions" have the same support as the original function, but are not guaranteed to be nonnegative; therefore one could reduce the degree of ambiguity by generating all possible "solutions" and then keeping only the nonnegative ones.[28]

For certain special types of one-dimensional functions, there is a high probability that the solution is unique. For a function having two separated intervals of support, being separated by an interval over which the function is zero, the solution usually is unique,[29,30] but only if the two intervals of support are sufficiently separated.[31] Another special type of function for which the solution is usually unique is one consisting of a summation of a number of delta-functions randomly distributed in space; for such functions, one does not need the iterative method—they can be reconstucted by a simple noniterative method involving the product of three

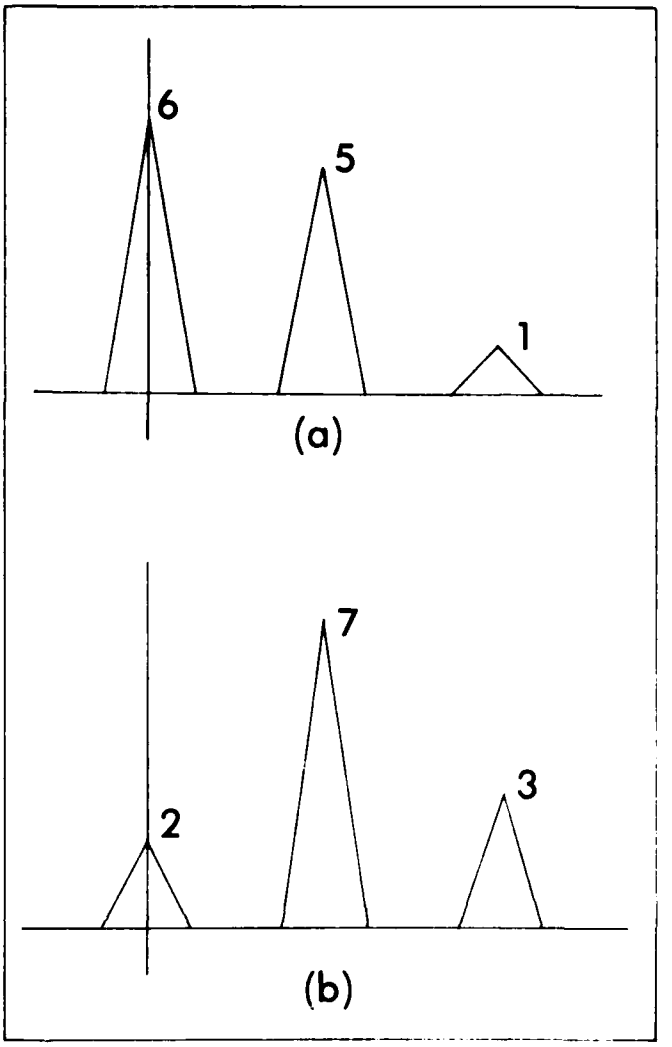


Fig. 6. Functions (a) and (b) having the same Fourier modulus.

translates of the autocorrelation function.[32]

In the event that multiple solutions do exist, it would not appear that the algorithm would be biased toward one over another, and one would expect the algorithm to converge to different solutions, depending on the initial input to the algorithm. For example, Fig. 6 shows two functions having the same Fourier modulus. In a computer experiment using the iterative reconstruction algorithm on the functions' Fourier modulus, it converged to one of the solutions in about half of the trials and converged to the other solution in the other half of the trials, depending on the random number sequences used as the initial input to the algorithm.

For the problem in two or more dimensions, it appears that the solution is usually unique. Considering sampled functions defined on a rectangular grid of points, Bruck and Sodin[33] showed that the existence of additional solutions is equivalent to the factorability of a polynomial representation of the Fourier transform. Since a polynomial of one variable of degree M can always be factored into M prime factors, there are 2^{M-1} solutions in the one-dimensional case. Once again, only some of the "solutions" may be nonnegative. On the other hand, polynomials of two or more variables having arbitrary coefficients are only rarely factorable; consequently, the two-dimensional problem is usually unique. Attempts have also been made to extend this concept to continuous, as opposed to discrete, functions.[34] Although it is always possible to make up examples in two dimensions that are not unique,[35] it appears to be

true that for two-dimensional functions drawn from the real world, the solution is usually unique. The general uniqueness of the two-dimensional case is indicated by experimental reconstruction results using the iterative algorithm.[36] Furthermore, noise in the Fourier modulus data has had the effect of adding noise to the reconstructed function rather than causing the algorithm to converge to a radically different solution.[37]

3.2.2. Astronomical reconstruction

The problem of reconstructing a two-dimensional nonnegative function from the modulus of its Fourier transform arises in astronomy. Due to atmospheric turbulence, the resolution attainable from large optical telescopes on earth is only about one second of arc, many times worse than the diffraction limit imposed by the diameter of the telescope aperture. For a five-meter telescope aperture, the diffraction-limited resolution would be about 0.02 seconds of arc—fifty times finer. Despite atmospheric turbulence, it is possible to measure the modulus of the Fourier transform of a space object out to the diffraction limit of the telescope using interferometric techniques.[38-41] The autocorrelation of the object can be computed from the Fourier modulus, allowing the diameter of the object to be determined. However, unless the Fourier transform phase is also measured, it was previously not possible to determine the object itself, except for some special cases. Previous attempts to solve this problem had not proven to be practical for complicated two-dimensional objects.

The problem of reconstructing an object from interferometer data can be solved by the iterative method.[42,36] The Fourier-domain constraint is that the Fourier modulus equal the Fourier modulus measured by an interferometer, and the function-domain constraint is that the object function be nonnegative. Figure 7 shows an example. Fig. 7(a) shows a computer-synthesized object used for the experiment—a sun-like disk having "solar flares" and bright and dark "sunspots." The modulus of its Fourier transform is shown in Fig. 7(b). Figure 7(c) shows a square of random numbers used as the initial input for the iterative algorithm. Figures 7(d), 7(e), and 7(f) show the reconstruction results after 20, 230, and 600 iterations, respectively. Figure 7(g) shows the initial input for a second trial, and the reconstruction results after 2 and 215 iterations are shown in Figs. 7(h) and 7(i), respectively. Comparing Figs. 7(f) and 7(i) with the original object in Fig. 7(a), one sees that for both trials, the reconstructed images match the original object very closely. Note that inverted solutions such as Fig. 7(f) are permitted for this problem since the modulus of the Fourier transform of f(-x) equals the modulus of the Fourier transform of f(x) for real-valued f(x). Other successful reconstruction experiments have been performed on data simulated to have the types of noise present in stellar speckle interferometry,[39] and it appears that under realistic levels of photon noise for fairly bright objects, diffraction-limited images can be reconstructed.[37] Initial expements have also been carried out on data from telescopes.[43]

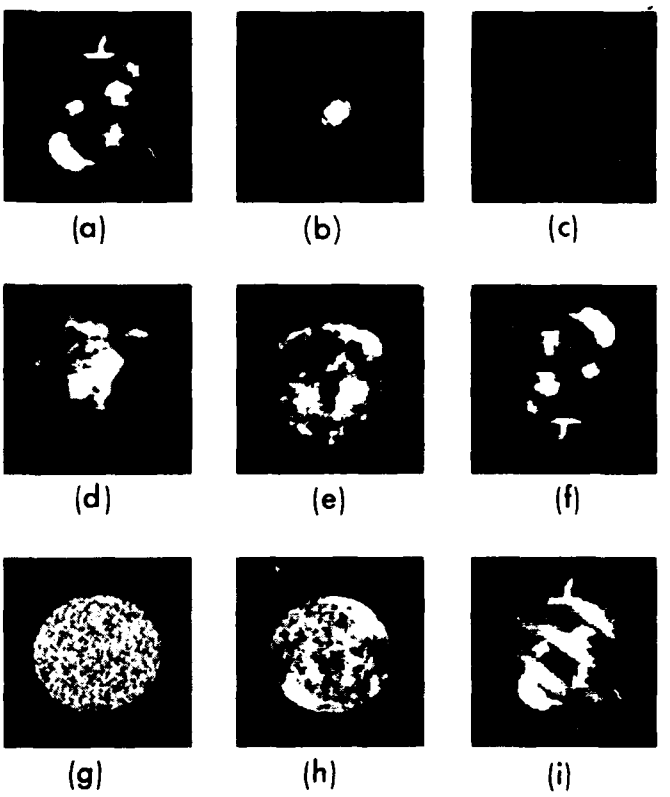


Fig. 7. Reconstruction of a nonnegative function from its Fourier modulus. (a) Test object; (b) modulus of its Fourier transform; (c) initial estimate of the object (first test); (d)-(f) reconstruction results —number of iterations: (d) 20, (e) 230, (f) 600; (g) initial estimate of the object (second test); (h)-(i) reconstruction results—number of iterations: (h) 2, (i) 215.

3.2.3. Pupil reconstruction and synthesis

Another case in which one may want to reconstruct a two-dimensional nonnegative function from its Fourier modulus is in pupil function determination. In a diffraction-limited optical system, the point-spread function is the squared Fourier modulus of the system's pupil function. Equivalently, the optical transfer function is the autocorrelation of the pupil function.[44] Given the point-spread function at a given location in an image plane, one could use the iterative algorithm to retrieve the corresponding pupil function, in a way that is mathematically equivalent to the astronomy problem. Turning this problem around, one could use the iterative algorithm to synthesize (design) a pupil function that would yield a given, desired point-spread function while possibly satisfying other desirable constraints as well.

3.3. Finite extent—measurement over part of an aperture

In a number of reconstruction problems, there is a function of known finite extent (or support) and one wishes to reconstruct the function with resolution appropriate to an aperture in the Fourier domain more complete than the one over which measurements were actually taken. In some cases, the desired aperture is simply larger than the aperture over which measurements were taken, and so one wishes to extrapolate the function's Fourier transform, i.e., to obtain superresolution of the function. In other cases, one has made measurements over a partially filled aperture, in which case one wishes to interpolate the Fourier transform of the function, and thereby obtain an improved impulse response in the function domain.

3.3.1. Extrapolation or superresolution

The error-reduction algorithm was first applied to the extrapolation (or superresolution) problem by Gerchberg.[45] Much has been written about the iterative algorithm, specifically the error-reduction algorithm, as it relates to this problem, including various ways of understanding the algorithm (see the end of Sec. 2) and proofs of convergence.[10,12,13,46-48] For this particular problem, the nature of the constraints makes it possible to implement the algorithm by a feedback optical processor[49,50] taking on the order of 10^{-9} seconds per iteration even for the two-dimensional case. Marks and Smith describe these matters in detail elsewhere in this volume.

3.3.2. Interpolation

In tomographic imaging systems, many projections of the object are measured, each projection yielding information about a slice through the Fourier transform of the object. When measurements over only a limited cone of angles are made, the effective aperture

in the Fourier domain has gaps, and the impulse response of the system is highly irregular. In applying the iterative algorithm to this problem,[51,52] the function-domain constraint is the finite extent and nonnegativity of the object, and the Fourier domain constraint is that the Fourier transform equal the measured Fourier transform over the measurement aperture.

A problem similar to the tomography problem arises in radio astronomy. The radio sky brightness map is a two-dimensional real, nonnegative function which is the Fourier transform of the complex visibility function. The visibility function is measured by radio interferometry, and in the case of long-baseline interferometry, the visibility function is measured only over a limited set of "tracks" in the Fourier domain, resulting in a partially-filled effective aperture. The error-reduction algorithm has been used to obtain improved maps by, in effect, interpolating the visibility function to fill in the area between the tracks.[53] For this problem, the constraints on the brightness map are that it be nonnegative and be zero outside the known field of view. In the visibility plane, the constraint is that the complex visiblity function equal the measured value within the area of the tracks.

3.4. Modulus—quantized values

As mentioned earlier in connection with spectrum shaping, in computer holography one may wish to encode the Fourier transform of an image as a computer-generated hologram, but some types of computer-generated holograms can encode only certain quantized complex values. The kinoform example discussed earlier is a special type of quantization. A more general example is the Lohmann hologram,[54] for which the modulus and phase of a complex sample are determined by the area and relative position, respectively, of an aperture within a sampling cell. The number of allowable quantized values is determined by the number of resolution elements, of the recording device used to fabricate the hologram, used to form one cell. For this synthesis problem, the function-domain constraint is that the modulus of the function equal the desired image modulus and the Fourier-domain constraint is that the complex Fourier coefficients fall on a prescribed set of quantized values. Experiments have shown that synthesizing such a Fourier transform pair is possible using the iterative algorithm.[55,7] For example, Fig. 8(a) shows a simulation of an image produced by a Lohmann hologram having only four modulus and four phase quantization levels when the image was random phase coded. Figure 8(b) shows the image after 13 iterations, a considerable improvement. This problem is one of a more general class of problems regarding the transmission of coded data.

3.5. Finite extent—phase

Finally, the iterative algorithm has been used to reconstruct the modulus of a band-limited signal from its phase.[56,57] Or, looking at it in another way, given that a function has finite extent and given the phase of its Fourier transform, reconstruct the modulus of its Fourier transform. For this application, it has been shown that for a wide class of conditions the solution is unique.[56] This application will be discussed further in Sec. 4.

4. ALGORITHM CONVERGENCE AND ACCELERATED ALGORITHMS

As mentioned in Sec. 2, the basic iterative algorithm depicted in Fig. 1, referred to as the error-reduction algorithm, has been shown to converge for some applications. In this section, the convergence is proven for all applications. In addition, modified algorithms that often converge much faster than the error-reduction algorithm are discussed.

4.1. Convergence of the error-reduction algorithm

For the error-reduction algorithm, the mean-squared error can be defined in general by Eq. (7) or Eq. (8). It is a normalized version of the integral over the square of the amount by which the computed function (or the computed Fourier transform) violates the constraints in the appropriate domain. When the mean-squared error is zero, then a Fourier transform pair has been found that satisfies all the constraints in both domains.

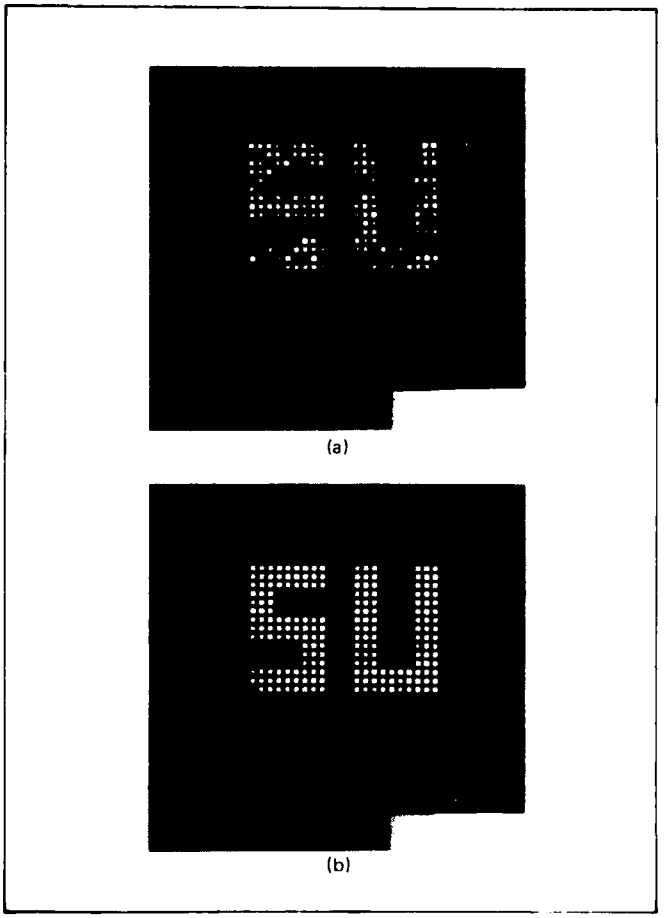


Fig. 8. Computer-simulated images from hologram with four magnitude and four phase quantized levels. (a) Object random phased coded; (b) after 13 iterations of the iterative method.

Consider again the steps in the error-reduction algorithm described in Sec. 2. The k^{th} iteration starts with an estimate $g_k(x)$ that satisfies the function-domain constraints. For any coordinate, x, the complex values that g(x) can have that satisfy the function-domain constraints form some set of points in phasor space. For example, if the modulus must equal $|f(x)|$, then the set of such points is a circle of radius $|f(x)|$ in phasor space; if the function must be nonnegative, then the set of such points is the half line on the nonnegative real axis. The function estimate $g_k(x)$ is Fourier transformed, yielding $G_k(u)$. The next step in the algorithm is to form $G'_k(u)$ by changing $G_k(u)$ by the smallest possible amount that allows it to satisfy the Fourier-domain constraints. $G'_k(u)$ is then inverse Fourier transformed, yielding $g'_k(x)$ in the function domain. In the final step, $g_{k+1}(x)$ is formed by changing $g'_k(x)$ by the smallest amount that allows it to satisfy the function-domain constraints. Now consider the unnormalized squared error, given by the numerators in Eqs. (7) and (8). In the Fourier domain, the unnormalized squared error at the k^{th} iteration is

$$e^2_{F\,k} = \int_{-\infty}^{\infty} |G_k(u) - G'_k(u)|^2 \, du \tag{14}$$

$$= \int_{-\infty}^{\infty} |g_k(x) - g_k'(x)|^2 \, dx ,$$

where the second line in this equation results from Parseval's theorem. The unnormalized squared error in the function domain at the k^{th} iteration is given by

$$e_{0k}^2 = \int_{-\infty}^{\infty} |g_{k+1}(x) - g_k'(x)|^2 \, dx . \tag{15}$$

Both $g_k(x)$ and $g_{k+1}(x)$ by definition satisfy the function-domain constraints. Also at any given coordinate x, $g_{k+1}(x)$ is the point in phasor space satisfying the function-domain constraints that is closest to $g_k'(x)$. Therefore, for all values of x,

$$|g_{k+1}(x) - g_k'(x)| \leq |g_k(x) - g_k'(x)| , \tag{16}$$

where equality holds only if $g_k(x)$ is just as close in phasor space to $g_k'(x)$ as $g_{k+1}(x)$ is. When there is a point in phasor space satisfying the constraints that is closer to $g_k'(x)$ than $g_k(x)$ is, then the left-hand side of the expression above is strictly less than the right-hand side. Therefore, combining Eqs. (14)–(16),

$$e_{0k}^2 \leq e_{Fk}^2 \tag{17}$$

for a given iteration. From the perfect symmetry of the error-reduction algorithm, as seen from Fig. 1, a similar result holds when one completes the iteration by satisfying the function-domain constraints, thereby forming $g_{k+1}(x)$, and continues the next iteration by Fourier transforming $g_{k+1}(x)$ and causing its transform to satisfy the Fourier-domain constraints. One then finds that

$$e_{F,k+1}^2 \leq e_{0k}^2 \leq e_{Fk}^2 . \tag{18}$$

Therefore, the unnormalized squared error can only decrease (or at least not increase) at each iteration. Since the normalized mean-squared error is simply proportional to the unnormalized squared error, a similar result holds for the errors defined by Eqs. (7) and (8).

While the error-reduction algorithm converges to a solution sufficiently fast for some applications, it is unbearably slow for others. In most cases, the error is reduced rapidly for the first few iterations, and then much more slowly for later iterations.

4.2. Input-output algorithms

Resulting from an investigation into the problem of the slow convergence of the error-reduction algorithm, a new and faster-converging algorithm was developed, the input-output algorithm.[55,58,7,36,42] The input-output algorithm differs from the error-reduction algorithm only in the function-domain operation. The first three operations—Fourier transforming g(x), satisfying Fourier domain constraints, and inverse Fourier transforming the result—are the same for both algorithms. Those three operations, if grouped together as shown in Fig. 9, can be considered as a nonlinear system with an input g(x) and an output g′(x). A property of this system is that its output is always a function having a Fourier transform that satisfies the Fourier-domain constraints. Therefore, if the output also satisfies the function-domain constraints, then all the constraints are satisfied and it is a solution to the problem. It is then necessary to determine how to manipulate the input in such a way as to force the output to satisfy the function-domain constraints.

For the error-reduction algorithm, the next input g(x) is chosen to be the current best estimate of the function satisfying the function-domain constraints. However, for the input-output algorithm, the input is not necessarily an estimate of the function or a modification of the output, nor does it have to satisfy the constraints; instead, it is viewed as the driving function for the next output. This viewpoint allows one a great deal of flexibility and inventiveness in selecting the next input and allows the invention of an algorithm that converges more rapidly to a solution. As will be seen later, the "input-output algorithm" actually comprises a few different algorithms, all of which are based on the input-output point-of-view.

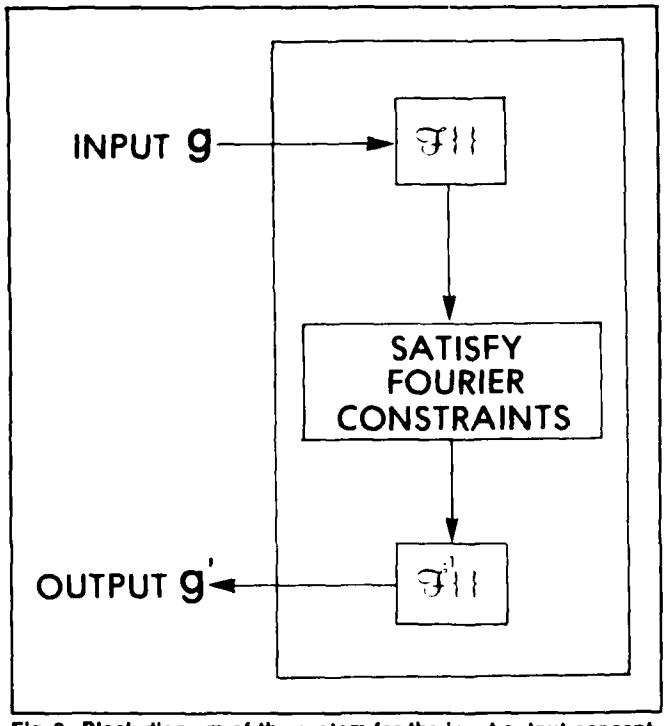


Fig. 9. Block diagram of the system for the input-output concept.

How the input should be changed in order to drive the output to satisfy the constraints depends on the particular problem at hand. The analysis given in the appendix for a specific application can be generalized as follows. Consider what happens when an arbitrary change is made in the input. Suppose that at the k^{th} iteration the input $g_k(x)$ results in the output $g_k'(x)$. Further, suppose that the input is then changed by adding $\Delta g(x)$:

$$g_{k+1}(x) = g_k(x) + \Delta g(x) . \tag{19}$$

Then one would expect the new output resulting from $g_{k+1}(x)$ to be of the form

$$g_{k+1}'(x) = g_k'(x) + \alpha \Delta g(x) + \text{additional noise.} \tag{20}$$

That is, the expected (or statistical mean) value of the change of the output, due to the change $\Delta g(x)$ of the input, is $\alpha \Delta g(x)$, a constant times the change of the input. The system shown in Fig. 9 is not linear; nevertheless, small changes of the input tend to result in similar changes of the output. The expected value of the change of the output can be predicted, but its actual value cannot be predicted since it has a non-zero variance. In the equation above, this lack of predictability is indicated by the "additional noise" term. The constant α depends on the statistics of $G_k(u)$ and $F(u)$ and on the Fourier-domain constraints.

If the output $g_k'(x)$ does not satisfy the function-domain constraints and if $g_k'(x) + \Delta g_d(x)$ does, then one might try to drive the

output to satisfy the constraints by changing the input in such a way as to cause the output to change by $\Delta g_d(x)$. According to the equation above, the change of the input that will, on the average, cause a change $\Delta g_d(x)$ of the output is

$$\Delta g(x) = \alpha^{-1} \Delta g_d(x) . \tag{21}$$

Thus a logical choice for the new input is

$$g_{k+1}(x) = g_k(x) + \beta \Delta g_d(x) , \tag{22}$$

where β is a constant ideally equal to α^{-1}, and where $\Delta g_d(x)$ is a function such that $g_k'(x) + \Delta g_d(x)$ satisfies the function-domain constraints. If α is unknown, then a value of β only approximately equal to α^{-1} will usually work nearly as well. The use of too small a value of β in Eq. (22) will only cause the algorithm to converge more slowly. The noise-like terms in Eq. (20) are kept to a minimum by minimizing $|\beta \Delta g_d(x)|$.

As mentioned earlier, for the input-output algorithm $g_k(x)$ is not necessarily an estimate of the function; it is instead the driving function for the next output. Therefore, it does not matter whether its Fourier transform, $G_k(u)$, satisfies the Fourier-domain constraints. Consequently, for the input-output algorithm, the mean-squared error, E_F^2, is unimportant; E_0^2 is the meaningful quality criterion. When computing E_0 for the input-output algorithm, the $g_{k+1}(x)$ that one should use in the integrand of Eq. (8) is the one determined by the error-reduction algorithm rather than the one computed by the input-output algorithm. That is, E_0 should still be a measure of the amount by which the output, $g_k'(x)$, violates the constraints.

Another interesting property of the system shown in Fig. 9 is that if an output $g'(x)$ is used as an input, then its output will be itself. Since the Fourier transform of $g'(x)$ already satisfies the Fourier-domain constraints, $g'(x)$ is unaffected as it goes through the system. Therefore, no matter what input actually resulted in the output $g'(x)$, the output $g'(x)$ can always be considered to have resulted from itself as an input. From this point of view, another logical choice for the new input is

$$g_{k+1}(x) = g_k'(x) + \beta \Delta g_d(x) \tag{23}$$

Note that if $\beta = 1$ in Eq. (23), then this version of the input-output algorithm reduces to the error-reduction algorithm. Since the optimum value of β is usually not unity, the error-reduction algorithm can be looked on as a suboptimal subset of one version of the more general input-output algorithm. Depending on the problem being solved, other variations in Eqs. (22) and (23) may be successful ways for choosing the next input.

In order to implement the input-output algorithm using Eq. (22) or (23), one chooses $\Delta g_d(x)$ according to the function-domain constraints. In general, a logical choice is the smallest value of $\Delta g_d(x)$ for which $g_k'(x) + \Delta g_d(x)$ satisfies the function-domain constraints. At those values of x for which $g_k'(x)$ already satisfies the function-domain constraints, one would set $\Delta g_d(x) = 0$. At those values of x for which $g_k'(x)$ violates the function-domain constraints, examples of logical choices of $\Delta g_d(x)$ for various applications are as follows. For the astronomy problem and other applications requiring the function to be nonnegative, choose $\Delta g_d(x) = -g_k'(x)$ where $g_k'(x)$ is negative. For applications requiring the function to be of finite extent, choose $\Delta g_d(x) = -g_k'(x)$ for x outside the known region of support. For applications requiring the function to have modulus equal to $|f(x)|$, choose

$$\Delta g_d(x) = |f(x)| \frac{g_k'(x)}{|g_k'(x)|} - g_k'(x) . \tag{24}$$

In addition to the values of $\Delta g_d(x)$ given above, there are other choices that are successful when used in Eqs. (22) and (23). Any $\Delta g_d(x)$ that moves $g'(x)$ in the general direction of satisfying the function-domain constraints will usually result in an algorithm that works; suboptimum choices of $\Delta g_d(x)$ and of β in Eq. (22) or Eq. (23) result in algorithms that converge less rapidly than the optimum. Two examples of other algorithms that converge more rapidly than the "logical" ones described in the preceding paragraph are as follows. For applications requiring the function to have modulus equal to $|f(x)|$, it was noticed that the difference in phase between $g_k'(x)$ and $g_k(x)$ tends to have the same sign as the change of phase of $g_k'(x)$ from one iteration to the next. In order to anticipate the direction that the phase is changing, one could choose a $\Delta g_d(x)$ that tends to rotate the phase angle of the new input toward that of the last output. That is, a good choice for the desired change of the output is

$$\Delta g_d(x) = \left[|f(x)| \frac{g_k'(x)}{|g_k'(x)|} - g_k'(x) \right] + \left[|f(x)| \frac{g_k'(x)}{|g_k'(x)|} - |f(x)| \frac{g_k(x)}{|g_k(x)|} \right] \tag{25}$$

in which the first component boosts (or shrinks) the magnitude of the output to match $|f(x)|$ and the second component rotates the phase angle of the input toward the phase angle of the output. For the astronomy problem, it was found that a particularly successful algorithm was to use Eq. (23) at those points where the constraints were satisfied and use Eq. (22) at those points where the constraints were violated, i.e.,

$$g_{k+1}(x) = \begin{cases} g_k'(x), & \text{where constraints satisfied} \\ g_k(x) - \beta g_k'(x), & \text{where constraints violated} \end{cases} \tag{26}$$

Furthermore, it was found that even faster convergence can be obtained by alternating between the above equation and the error-reduction algorithm every few iterations.

Unlike the error-reduction algorithm, the input-output algorithm is not guaranteed to converge; in fact the error may even increase for some of the iterations. However, the input-output algorithm is much less prone to stagnation and therefore in practice converges much faster than the error-reduction algorithm. In some instances during the input-output iterations, E_0 may even increase although the visual appearance of the image improves. This behavior, which is poorly understood, is described further in Ref. 59.

From the paragraphs above, it is seen that the "input-output algorithm" is really a family of algorithms. The input-output approach is one that can lead to a number of different algorithms based on the manner in which the nonlinear system of Fig. 9 behaves. One would hope that the principles of control theory and possibly other disciplines could be used to shed further light on this system and help to arrive at algorithms with still more rapid convergence.

It should also be noted that, unlike the error-reduction algorithm, the input-output algorithm does not treat the two domains in a symmetric manner. By reversing the roles of the two domains, one can arrive at a different and possibly more advantageous algorithm.

4.3. Relaxation-parameter algorithm

A second method of improved convergence is the use of a relaxation parameter. In solving the problem of reconstructing the magnitude of a band-limited function from its phase (or, equivalently, reconstructing a function of finite extent from the

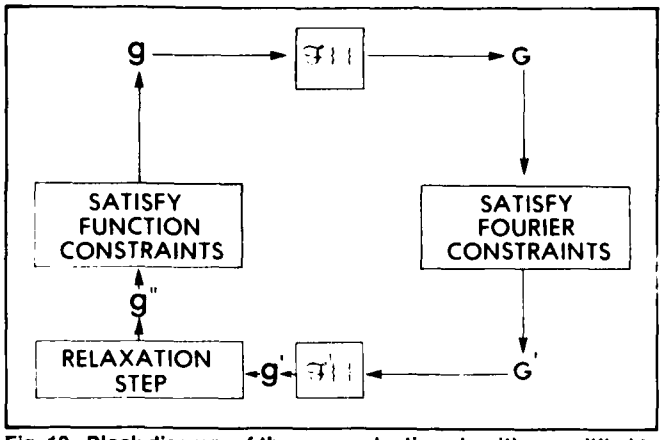


Fig. 10. Block diagram of the error-reduction algorithm modified to include a relaxation step.

phase of its Fourier transform), Oppenheim, Hayes, and Lim[57] modified the error-reduction algorithm (Fig. 1) by adding a relaxation step, as shown in Fig. 10. Here the band-limited function is taken to be in the Fourier domain. The function g(x) then must be of finite extent according to the bandwidth of the Fourier-domain function. In the relaxation step, $g_k''(x)$ is formed from $g_k'(x)$ according to

$$g_k''(x) = (1 - \eta_k)g_{k-1}''(x) + \eta_k g_k'(x), \tag{27}$$

and then the new estimate $g_{k+1}(x)$ is formed from $g_k''(x)$ by making it satisfy the function-domain constraints. The parameter η_k, which is a constant that may vary from one iteration to the next, is the relaxation parameter. For $\eta_k = 1$, $g_k''(x) = g_k'(x)$ and this reduces to the error-reduction approach. For $\eta_k = 0$, $g_k''(x) = g_{k-1}''(x)$, that is, the result from the previous iteration is used. Other values of η_k give a linear combination of $g_{k-1}''(x)$ and $g_k'(x)$. For the reconstruction of a function of finite extent from the phase of its Fourier transform or from a segment of its Fourier transform (i.e., the superresolution problem), if $g_1'(x)$ and $g_2'(x)$ both satisfy the Fourier-domain constraint, then the linear combination $\eta g_1'(x) + (1 - \eta)g_2'(x)$ also satisfies the constraint in the Fourier domain. It follows from this that $g_k''(x)$ given by Eq. (27) also satisfies the Fourier-domain constraint. In those cases, it can be shown that the algorithm converges for $0 < \eta_k \leq 1$. However, for other sets of constraints, for example, given the modulus of the Fourier transform, $g_k''(x)$ given by the equation above does not generally satisfy the Fourier-domain constraints and so the relaxation method does not strictly apply.

The optimum value of η_k can be determined as follows. Define the function-domain squared error after the relaxation step as

$$e_0^2 = \int_\gamma |g_k''(x)|^2 \, dx, \tag{28}$$

where the region of integration, γ, is the region over which the function is known to be zero. Setting equal to zero the derivative of e_0^2 with respect to η_k, and solving for η_k, one finds the optimum value of η_k to be given by

$$\eta_k = \frac{-\mathrm{Re}\left\{\int_\gamma g_{k-1}''(x)[g_k'(x) - g_{k-1}''(x)]^* \, dx\right\}}{\int_\gamma |g_k'(x) - g_{k-1}''(x)|^2 \, dx}. \tag{29}$$

The computation of the relaxation parameter by Eq. (29) takes much less time than the computation of one (fast) Fourier transform, and so it does not significantly increase the total computation time of a single iteration.

Use of the relaxation step for the problem of reconstructing a band-limited function from its phase resulted in an order of magnitude improvement in the speed of convergence of the algorithm over that of the error-reduction algorithm.[57]

The relaxation step described above incorporates the optimum combination of the current output with the previous output. It is also possible to extend this concept to include a number of previous outputs,[57] which may result in still more rapid convergence.

It should be noted that the majority of the work referenced in Sec. 3 made use of only the error-reduction algorithm. Improved speed of convergence could be expected if one of the two accelerated algorithms discussed above were employed.

5. SUMMARY AND COMMENTS

The iterative error-reduction algorithm, an extension of the Gerchberg-Saxton algorithm to include various types of constraints, has been found to be capable of solving a wide range of difficult problems in optics and other fields. It can be applied to the reconstruction of a function (an object, wavefront, signal, etc.) when only partial information is available in each of two domains, or to the synthesis of a function (wavefront, signal, etc.) having desired properties in each of two domains. The iterative algorithm is reasonably fast for most applications, since the major computational burden, two Fourier transforms per iteration, can be accomplished using the fast Fourier transform (FFT) algorithm. The iterative algorithm has been shown to outperform alternative methods of solving these classes of problems both because of its speed and its tolerance of noise.[4,9] For some applications, a large number of iterations is required for convergence of the error-reduction algorithm. This situation can be remedied by using an algorithm with accelerated convergence, such as the input-output algorithm or an algorithm employing a relaxation step.

The iterative algorithm has been in use for only a few years, yet it has already found numerous applications; and methods of improving the algorithm have been devised. Nevertheless, it is safe to predict that it will be used in the future to solve new problems not discussed here, and it is hoped that further improvements of the algorithm will be discovered.

POSTSCRIPT

As this book goes to print, further developments relating to the iterative algorithm are occurring at a rapid pace. It has been uncovered that an algorithm equivalent to Gerchberg's[45] error-reduction algorithm for extrapolation was proposed by Ville[60] in 1956, although approached from a different point of view. Relationships between the error-reduction algorithm and gradient search methods have been discovered[59,61,62] and uncovered.[63] And further work on various applications is being reported.[64-83]

APPENDIX: ANALYSIS OF THE INPUT-OUTPUT SYSTEM

Consider the synthesis problem for kinoforms, for which the Fourier modulus is set equal to a constant. Suppose that the input g(x) to a kinoform system results in the output g'(x). The kinoform has a transmittance $G'(u) = K \exp[i\phi(u)]$, where $\phi(u)$ is the phase of $G(u) = |G(u)| \exp[i\phi(u)] = \mathcal{F}[g(x)]$, and K is a constant. The resulting image is $g'(x) = \mathcal{F}^{-1}[G'(u)]$. Now consider what happens when a change $\Delta g(x)$ is made in the input. As illustrated in the phasor diagrams in Fig. A1, the change $\Delta g(x)$ of the input causes a change $\Delta G(u)$ of its Fourier transform, which causes a change $\Delta G'(u)$ of the kinoform and a corresponding change $\Delta g'(x) = \mathcal{F}^{-1}[\Delta G'(u)]$ of the output image. The goal here is to determine the relationship between the change $\Delta g'(x)$ of the output and the change $\Delta g(x)$ of the input. Figure A2 shows the relationship be-

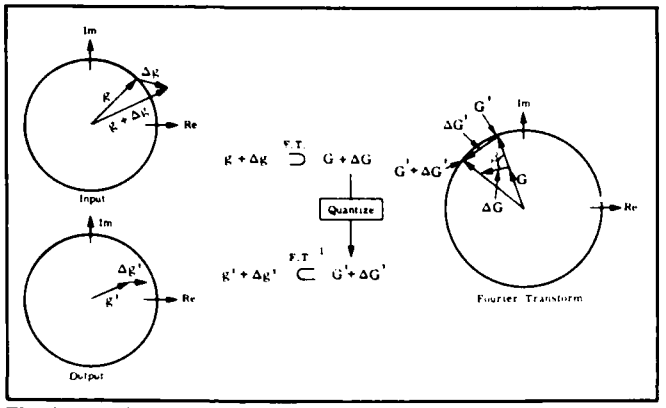


Fig. A1. A change Δg of the input results in a change ΔG′ of the kinoform and a change of Δg′ of the output.

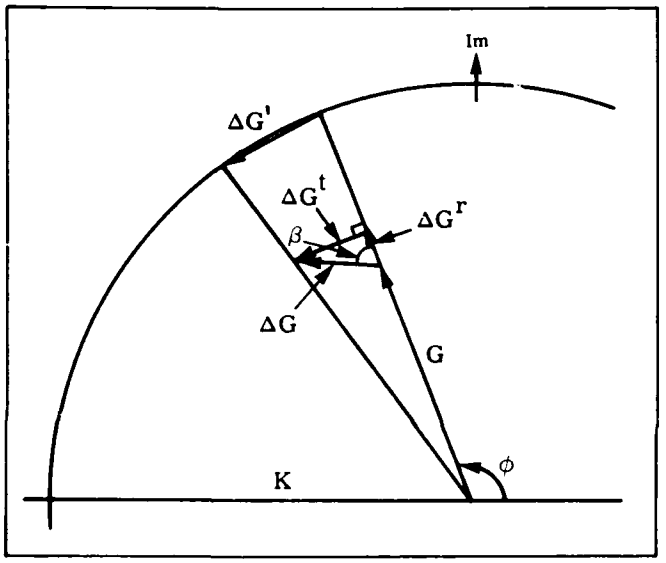


Fig. A2. Relationship between ΔG′, the change of the kinoform, and two components of ΔG, the Fourier transform of the change of the input.

tween $\Delta G'(u)$ and two orthogonal components of $\Delta G(u)$. By similar triangles, for $\Delta G' << |G|$,

$$\Delta G'(u) \simeq \Delta G^t(u) \frac{K}{|G(u)|}, \tag{A1}$$

where the two orthogonal components of $\Delta G(u)$ are

$$\Delta G^r(u) = \Delta G(u) \cos\beta(u)\, e^{i\phi(u)} \tag{A2}$$

parallel to G(u), and

$$\Delta G^t(u) = \Delta G(u) \sin\beta(u)\, e^{i[\phi(u) + \pi/2]} \tag{A3}$$

orthogonal to G(u); and

$$\Delta G(u) = \Delta G^r(u) + \Delta G^t(u) = \Delta G(u)\, e^{i[\phi(u) + \beta(u)]}, \tag{A4}$$

where $\beta(u)$ is the angle between $\Delta G(u)$ and G(u). Only one of the two orthogonal components of $\Delta G(u)$, namely $\Delta G^t(u)$, contributes to $\Delta G'(u)$.

In order to compute the expected change of the output, $E[\Delta g'(x)]$, treat the phase angles $\beta(u)$ and the magnitudes $|G(u)|$ as random variables. Inserting $|\Delta G(u)|$ from Eq. (A4) into Eq. (A3), one obtains

$$\Delta G^t(u) = \Delta G(u)\, e^{-i[\phi(u) + \beta(u)]} \sin\beta(u)\, e^{i\phi(u)}\, e^{i\pi/2}$$

$$= \Delta G(u)[\sin^2\beta(u) + i \sin\beta(u)\cos\beta(u)]. \tag{A5}$$

For $\beta(u)$ uniformly distributed over $[0, 2\pi]$,[19] the expected value of $\Delta G^t(u)$ is

$$E[\Delta G^t(u)] = \Delta G(u)\left(\frac{1}{2} + i\cdot 0\right) = \frac{1}{2}\Delta G(u). \tag{A6}$$

Therefore, the expected value of the change of the output is, using Eqs. (A1) and (A6) and assuming that the magnitudes $|G(u)|$ are identically distributed random variables[19] independent of $\beta(u)$,

$$E[\Delta g'(x)] = E\left[\mathcal{F}(\Delta G')\right]$$

$$\simeq \mathcal{F}[E(\Delta G')] \simeq \mathcal{F}\left[E(\Delta G^t)\, E\left(\frac{K}{|G|}\right)\right]$$

$$\simeq \mathcal{F}\left[\frac{1}{2}\Delta G(u)\right] E\left(\frac{K}{|G|}\right) = \frac{1}{2}\Delta g(x) E\left(\frac{K}{|G|}\right). \tag{A7}$$

That is, the expected change of the output is α times the change of the input, giving us the second term in Eq. (20), where $\alpha = (1/2)E(K/|G|)$. After a few iterations, G(u) will not differ greatly from K; then $\alpha \simeq 1/2$.

Similarly, the variance of the change of the output can be shown to be[58]

$$E[\Delta g'(x)^2] - E[\Delta g'(x)]^2$$

$$\simeq \frac{1}{4}\left\{2E\left(\frac{K^2}{|G|^2}\right) - \left[E\left(\frac{K}{|G|}\right)\right]^2\right\} \cdot \frac{1}{A}\int_{-\infty}^{\infty} \Delta g(x')^2\, dx', \tag{A8}$$

where A is the area of the image. That is, the variance of the change of the output $\Delta g'(x)$ at any given x is proportional to the integrated squared change of the entire input. The predictability of $\Delta g'(x)$, and the degree of control with which one can manipulate it, decreases as one makes larger changes of the input. The difference between the actual change of the output and the expected change of the output given by Eq. (A7) is what is meant by the additional noise term in Eq. (20). If, after a few iterations, $G(u) \simeq K$, then in Eq. (A8) the factor $(1/4)\{2E(K^2/|G|^2) - [E(K/|G|)]^2\} = 1/4$.

Equations (A7) and (A8) are a justification for the input-output concept: small changes of the input result in similar changes of the output, and so the output can be driven to satisfy the constraints by appropriate changes of the input, as in Eqs. (22) and (23).

ACKNOWLEDGMENT

The author gratefully acknowledges the support of the U.S. Air Force Office of Scientific Research.

REFERENCES

1. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik 35, 237 (1972).
2. P. M. Hirsch, J. A. Jordan, Jr., and L. B. Lesem, "Method of making an object-dependent diffuser," U.S. Patent No. 3,619,022 (Nov. 9, 1971; filed Sept. 17, 1970).
3. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," Appl. Opt. 12, 2328 (1973).
4. W. O. Saxton, *Computer Techniques for Image Processing in Electron Microscopy* (Academic Press, New York, 1978).
5. W. T. Cochran, J. W. Cooley et al., "What is the fast Fourier transform?" Proc. IEEE 55, 1664 (1967).
6. B. Liu and N. C. Gallagher, "Convergence of a spectrum shaping algorithm," Appl. Op 13, 2470 (1974).
7. J. R. Fienup, "Iterative method applied to image reconstruction and to computer-generated holograms," Opt. Eng. 19(3), 297 (1980).
8. D. L. Misell, "A method for the solution of the phase problem in electron microscopy," J. Phys. D.: Appl. Phys. 6, L6-L9 (1973); D. L. Misell, "An examination of an iterative method for the solution of the phase problem in optics and electron optics," J. Phys. D.: Appl. Phys. 6, 2200 (1973).
9. R. Boucher, "Convergence of algorithms for phase retrieval from two intensity distributions," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 130 (1980).
10. R. W. Schafer, R. M. Mersereau, and M. A. Richards, "Constrained iterative restoration algorithms," Proc. IEEE 69, 432 (1981).
11. G. Dahlquist and A. Björck (translated by N. Anderson), *Numerical Methods* (Prentice-Hall, Englewood Cliffs, N.J., 1974) pp. 2-4.
12. D. C. Youla, "Generalized image restoration by method of alternating orthogonal projections," IEEE Trans. Circuits and Systems CAS-25, 694 (1978).
13. M. S. Sabri and W. Steenaart, "An approach to band-limited signal extrapolation: The extrapolation matrix," IEEE Trans. Circuits and Systems CAS-25, 74 (1978).
14. A. M. J. Huiser, P. Van Toorn, and H. A. Ferwerda, "On the problem of phase retrieval in electron microscopy from image and diffraction pattern I-IV," Optik 47, 123 (1977); J. Gassmann, "Optimal iterative phase retrieval from image and diffraction intensities," Optik 48, 347 (1977).
15. A. M. J. Huiser, A. J. J. Drenth, and H. A. Ferwerda, "On phase retrieval in electron microscopy from image and diffraction pattern," Optik 45, 303 (1976); A. M. J. Huiser and H. A. Ferwerda, "On the problem of phase retrieval in electron microscopy from image and diffraction pattern II: on the uniqueness and stability," Optik 46, 407 (1976); A. J. Devaney and R. Chidlaw, "On the uniqueness question in the problem of phase retrieval from intensity measurements," J. Opt. Soc. Am. 68, 1352 (1978).
16. T. S. Huang, "Digital holography," Proc. IEEE 59, 1335 (1971); W.-H. Lee, "Computer-generated holograms: techniques and applications," in E. Wolf, ed., *Progress in Optics*, Vol. 16 (North-Holland, 1978) pp. 121-232; W. J. Dallas, "Computer-generated holograms," in B. R. Frieden, ed., *The Computer in Optical Research* (Springer-Verlag, N.Y., 1980), Chapter 6.
17. L. B. Lesem, P. M. Hirsch, and J. A. Jordan, Jr., "The kinoform: A new wavefront reconstruction device," IBM J. Res. Develop. 13, 150 (1969).
18. H. Akahori, "Comparison of deterministic phase coding with random phase coding in terms of dynamic range," Appl. Opt. 12, 2336 (1973).
19. R. S. Powers and J. W. Goodman, "Error rates in computer-generated holographic memories," Appl. Opt. 14, 1690 (1975).
20. W.-H. Lee, "Method for converting a Gaussian laser beam into a uniform beam," Opt. Commun. 36, 469 (1981).
21. R. A. Gonsalves, "Phase retrieval from modulus data," J. Opt. Soc. Am. 66, 961 (1976).
22. J. Maeda and K. Murata, "Retrieval of wave aberration from point spread function or optical transfer function data," Appl. Opt. 20, 274 (1981).
23. R. N. Bracewell, *The Fourier Transform and Its Applications*, 2nd Edition (McGraw-Hill, New York, 1978).
24. E. Wolf, "Is a complete determination of the energy spectrum of light possible from measurements of the degree of coherence?" Proc. Phys. Soc. (London) 80, 1269 (1962).
25. G. H. Stout and L. H. Jensen, *X-Ray Structure Determination* (Macmillan, London, 1968).
26. A. Walther, "The question of phase retrieval in optics," Optica Acta 10, 41 (1963).
27. E. M. Hofstetter, "Construction of time-limited functions with specified autocorrelation functions," IEEE Trans. Info. Theory IT-10, 119 (1964).
28. R. H. T. Bates, "Contributions to the theory of intensity interferometry," Mon. Not. R. Astr. Soc. 142, 413 (1969).
29. A. H. Greenaway, "Proposal for phase recovery from a single intensity distribution," Opt. Lett. 1, 10 (1977).
30. R. H. T. Bates, "Fringe visibility intensities may uniquely define brightness distributions," Astron. and Astrophys. 70, 127-129 (1978).
31. T. R. Crimmins and J. R. Fienup, "Ambiguity of phase retrieval for functions with disconnected support," J. Opt. Soc. Am. 71, 1026 (1981).
32. J. R. Fienup, T. R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation," J. Opt. Soc. Am. 72, 610 (1982).
33. Yu. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. 30, 304 (1979).
34. W. Lawton, "A numerical algorithm for 2-D wavefront reconstruction from intensity measurements in a single plane," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 94 (1980).
35. A. M. J. Huiser and P. Van Toorn, "Ambiguity of the phase-reconstruction problem," Opt. Lett. 5, 499 (1980).
36. J. R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. 18, 529 (1979).
37. G. B. Feldkamp and J. R. Fienup, "Noise properties of images reconstructed from Fourier modulus," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 84 (1980).
38. A. Labeyrie, "Attainment of diffraction limited resolution in large telescopes by Fourier analysing speckle patterns in star images," Astron. and Astrophys. 6, 85 (1970).
39. D. Y. Gezari, A. Labeyrie, and R. V. Stachnik, "Speckle interferometry: diffraction-limited measurements of nine stars with the 200-inch telescope," Astrophys. J. Lett. 173, L1-L5 (1972).
40. D. G. Currie, S. L. Knapp, and K. M. Liewer, "Four stellar-diameter measurements by a new technique: amplitude interferometry," Astrophys. J. 187, 131 (1974).
41. R. Hanbury Brown and R. Q. Twiss, "Correlation between photons in two coherent beams of light," Nature 177, 27 (1956).
42. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. 3, 27 (1978).
43. J. R. Fienup and G. B. Feldkamp, "Astronomical imaging by processing stellar speckle interferometry data," in *Applications of Speckle Phenomena*, W. H. Carter, ed., Proc. SPIE 243, 95 (1980).
44. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, San Francisco, 1968).
45. R. W. Gerchberg, "Super-resolution through error energy reduction," Optica Acta 21, 709 (1974).
46. A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," IEEE Trans. Circuits and Systems CAS-22, 735 (1975).
47. J. A. Cadzow, "An extrapolation procedure for band-limited signals," *IEEE Trans. Acoust., Speech, Signal Processing* ASSP-27, 4 (1978).
48. C. K. Rushforth and R. L. Frost, "Comparison of some algorithms for reconstructing space-limited images," J. Opt. Soc. Am. 70, 1539 (1980).
49. R. J. Marks II, "Coherent optical extrapolation of 2-D band-limited signals: processor theory," Appl. Opt. 19, 1670 (1980).
50. R. J. Marks II and David K. Smith, "Iterative coherent processor for band-limited signal extrapolation," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 106 (1980).
51. K.-C. Tam and V. Perez-Mendez, "Limited-angle 3-D reconstructions using Fourier transform iterations and Radon transform iterations," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 142 (1980).
52. T. Sato, S. J. Norton et al., "Tomographic image reconstruction from limited projections using iterative revisions in image and transform spaces," Appl. Opt. 20, 395 (1981).
53. A. E. E. Rogers, "Method of using closure phases in radio aperture synthesis," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 10 (1980).
54. B. R. Brown and A. W. Lohmann, "Computer-generated binary holograms," IBM J. Res. Develop. 13, 160 (1969); A. W. Lohmann and D. P. Paris, "Binary Fraunhofer holograms, generated by computer," Appl. Opt. 6, 1739 (1967).
55. J. R. Fienup, "Reduction of quantization noise in kinoforms and computer-generated holograms," J. Opt. Soc. Am. 64, 1395 (1974) (Abstract).
56. M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," IEEE Trans. Acoust., Speech, Signal Processing ASSP-28, 672 (1980).
57. A. V. Oppenheim, M. H. Hayes, and J. S. Lim, "Iterative procedure for signal reconstruction from phase," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE 231, 121 (1980).
58. J. R. Fienup, "Improved synthesis and computational methods for computer-generated holograms," Ph.D. Thesis, Stanford University, May 1975 (University Microfilms No. 75-25523), Chapter 5.
59. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. 21, 2758 (1982).
60. J.-A. Ville, "Sur le prolongement des signaux a spectre borne," Cables et Transmission 1, 44 (1956).

61. H. Maitre, "Iterative superresolution: some new fast methods," Opt. Acta 28, 973 (1981).
62. A. K. Jain and S. Ranganath, "Extrapolation algorithms for discrete signals with application in spectral estimation," IEEE Trans. Acoustics, Speech, Signal Processing ASSP-29 830 (1981).
63. M. T. Manry and J. K. Aggarwal, "The design of multi-dimensional FIR digital filters by phase correction," IEEE Trans. Circuits and Systems CAS-23, 185 (1976).
64. J. N. Mait and W. T. Rhodes, "Iterative design of pupil functions for bipolar incoherent spatial filtering," in *Processing of Images and Data from Optical Sensors*, W. H. Carter, ed., Proc. SPIE 292, 66 (1981).
65. J. G. Walker, "Object reconstruction from turbulence-degraded images," Opt. Acta 28, 1017 (1981).
66. H. Stark, D. Cahana, and H. Webb, "Restoration of arbitrary finite-energy optical objects from limited spatial and spectral information," J. Opt. Soc. Am. 71, 635 (1981).
67. K. C. Tam and V. Perez-Mendez, "Tomographic imaging with limited-angle input," J. Opt. Soc. Am. 71, 582 (1981).
68. T. F. Quatieri, Jr., and A. V. Oppenheim, "Iterative technique for minimum phase signal reconstruction from phase or magnitude," IEEE Trans. Acoustics, Speech, Signal Processing ASSP-29, 1187 (1981).
69. A. V. Oppenheim, "The importance of phase in signals," Proc. IEEE 69, 529 (1981).
70. L. S. Taylor, "The phase retrieval problem," IEEE Trans. Antennas Propagation AP-3, 386 (1981).
71. R. A. Gonsalves, "Phase retrieval and diversity in adaptive optics," Opt. Eng. 21, 829 (1982).
72. T. Sato, K. Saski, Y. Nakamura, M. Linzer, and S. J. Norton, "Tomographic image reconstruction from limited projections using coherent feedback," Appl. Opt. 20, 3073 (1981).
73. D. Cahana and H. Stark, "Bandlimited image extrapolation with faster convergence," Appl. Opt. 20, 2780 (1981).
74. J. R. Fienup, "Image reconstruction for stellar interferometry," in *Current Trends in Optics*, F. T. Arecchi and F. R. Aussenegg, eds (Taylor and Francis, London, 1981) pp. 95-102.
75. V. T. Tom, T. F. Quatieri, M. H. Hayes and J. H. McClellan, "Convergence of iterative nonexpansive signal reconstruction algorithms," IEEE Trans. Acoustics, Speech, Signal Processing ASSP 29, 1052 (1981).
76. J. S. Lim and N. A. Malik, "A new algorithm for two-dimensional maximum entropy power spectrum estimation," IEEE Trans Acoustics, Speech, Signal Processing ASSP-29, 401 (1981).
77. A. Lent, "An iterative method for the extrapolation of band limited functions," J. Math. Analysis and Applications 83, 554 (1981).
78. W. D. Montgomery, "Optical applications of von Neumann's alternating-projection theorem," Opt. Lett. 7, 1 (1982).
79. W. D. Montgomery, "Restoration of images possessing a finite Fourier series," Opt. Lett. 7, 54 (1982).
80. N. C. Gallagher and D. W. Sweeney, "Infrared holographic optical elements with applications to laser material processing," IEEE J Quantum Electronics QE-15, 1369 (1979).
81. F. A. Grunbaum, "A study of Fourier space methods for limited angle image reconstruction," Numer. Funct. Anal. and Optiomiz 2, 31 (1980).
82. I. Kadar, "A robustized vector recursive stabilizer algorithm for image restoration," Information and Control 44, 320 (1980).
83. R. Goutte, R. Prost, and A. Georges, "Déconvolution numérique avec prolongement spectral applications aux signaux et aux images," Analysis 8, 6 (1980).

Appendix B

RECONSTRUCTION OF OBJECTS HAVING LATENT REFERENCE POINTS

J.R. Fienup

Reprinted from the Journal of the Optical Society of America 73, 1421-1426 (November 1983).

Reprinted from Journal of the Optical Society of America, Vol. 73, page 1421, November 1983

Reconstruction of objects having latent reference points

J. R. Fienup

Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107

Received March 18, 1983; revised manuscript received July 14, 1983

A simple recursive algorithm is proposed for reconstructing certain classes of two-dimensional objects from their autocorrelation functions (or equivalently from the modulus of their Fourier transforms—the phase-retrieval problem). The solution is shown to be unique in some cases. The objects contain reference points not satisfying the holography condition but satisfying weaker conditions. Included are objects described by Fiddy *et al.* [Opt. Lett. 8, 96 (1983)] satisfying Eisenstein's thorem.

INTRODUCTION

In a number of disciplines, including astronomy, x-ray crystallography, electron microscopy, and wave-front sensing, one encounters the phase-retrieval problem. One wishes to reconstruct $f(m, n)$, an object function, from $|F(p, q)|$, the modulus of its Fourier transform, where

$$F(p,q) = |F(p,q)|\exp[i\psi(p,q)] = \mathcal{F}[f(m,n)]$$
$$= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\exp[-i2\pi(mp/M + nq/N)], \quad (1)$$

where $m, p = 0, 1, \ldots, M-1$ and $n, q = 0, 1, \ldots, N-1$. The discrete transform is employed here since in practice one deals with sampled data in a computer. The problem of reconstructing the object from its Fourier modulus is equivalent to reconstructing the Fourier phase, $\psi(p, q)$, from the Fourier modulus; since once one has the phase as well as the modulus, one can easily compute $f(m, n)$ by the inverse (discrete) Fourier transform. $r_f(m, n)$, the (aperiodic) autocorrelation of $f(m, n)$, is given by[1]

$$r_f(m,n) = \sum_{j=0}^{M-1}\sum_{k=0}^{N-1} f(j,k)f^*(j-m, k-n) \quad (2)$$

$$= \mathcal{F}^{-1}[|F(p,q)|^2], \quad (3)$$

where the asterisk denotes complex conjugate. Note that the autocorrelation is Hermitian: $r_f(-m, -n) = r_f^*(m, n)$. Note also that in order to avoid aliasing during the computation of $|F(p, q)|^2$, it is necessary to have $f(m, n) = 0$ for $M/2 \le m \le M - 1$ and for $N/2 \le n \le N - 1$; this will be assumed throughout this paper. Then there is no difference between the periodic (cyclic) and aperiodic autocorrelation. (For x-ray crystallography this is usually not the case, and the results of this paper do not apply.) Since the autocorrelation function is easily computed from the Fourier modulus by Eq. (3), the phase-retrieval problem is equivalent to reconstructing an object from its autocorrelation function.

Several phase-retrieval algorithms have been proposed, all of them requiring some additional measurements or constraints on the solution. Examples include a reference point at least one object diameter from the object[2] (giving rise to the holography condition[3]), a second intensity measurement in another plane[4,5] (in electron microscopy or wave-front sensing), nonnegativity and limited spatial extent[6-8] (in astronomy), atomic models[9] (in x-ray crystallography), and objects consisting of collections of points having nonredundant spacings.[10]

Here it is pertinent to review the case of holography. Suppose that $f(m, n)$ consists of an object of interest, $g(m, n)$, plus an unresolved (delta-function-like) point, referred to as the reference point, i.e.,

$$f(m,n) = A\delta(m - m_0, n - n_0) + g(m,n), \quad (4)$$

where $\delta(m, n)$ is a two-dimensional (2-D) Kronecker delta function. Then the autocorrelation can be written as the sum of four terms,

$$r_f(m,n) = |A|^2\delta(m,n) + r_g(m,n) + Ag^*(m_0 - m, n_0 - n) + A^*g(m + m_0, n + n_0), \quad (5)$$

the final term of which is the cross-correlation of the reference point with the object of interest and is simply proportional to a translate of the object of interest. If the distance from the reference point to the object of interest exceeds the diameter of the object of interest, then the fourth term in Eq. (5) is nonoverlapping with the other terms, and the object of interest is reconstructed by simple inspection of the autocorrelation. Then the holography condition is satisfied.[2,3] If the amplitude and position of the reference point are unknown (except that the reference point satisfies the holography condition), then the object can be reconstructed only to within a complex factor A^* and to within a translation, and there would be a twofold ambiguity as to whether the object is given by the fourth term or the third term (the conjugate image) of Eq. (5).

In this paper we describe an algorithm for reconstructing certain objects having reference points that do not satisfy the holography condition. For these cases the reference points may be referred to as latent reference points, because they do not immediately yield the object as would a holographic reference point; rather, a degree of development is required before their usefulness emerges.

In Section 2 the question of the uniqueness of the solution is reviewed. In Section 3 the new reconstruction algorithm is described as it is applied to three different classes of objects. Additional comments on the reconstruction algorithm are included in Section 4.

0030-3941/83/111421-06$01.00

2. UNIQUENESS OF THE SOLUTION

When one measures only the Fourier modulus, then the uniqueness of the solution is a central question. One of course always has the twofold (180° rotated or conjugate image) ambiguity since $|\mathcal{F}[f(m,n)]| = |\mathcal{F}[f^*(-m,-n)]|$; and translations of $f(m,n)$ and the multiplication of $f(m,n)$ by a constant phase factor $\exp(i\theta)$ (where θ is a real constant) also have no effect on $|F(p,q)|$. If these are the only ambiguities, then we consider the solution of the phase-retrieval problem to be unique.

Bruck and Sodin[11] considered objects consisting of a rectangular grid of delta functions having various complex amplitudes (or equivalently, a 2-D sequence), which have Fourier transforms that can be expressed as polynomials. These are the types of objects assumed by Eqs. (1) and (2), and we refer to such objects as sampled objects. They showed that, for sampled objects, a lack of uniqueness of the solution to the phase-retrieval problem is equivalent to the factorability of the polynomial, and therefore one-dimensional (1-D) objects of length L have a 2^{L-1}-fold ambiguity.[11] This result corresponds to the analogous theory for 1-D continuous functions.[12] On the other hand, polynomials of two (or more) variables are known to be only rarely factorable (i.e., they are usually irreducible). Consequently, for 2-D sampled objects the solution to the phase-retrieval problem is usually unique. An analogous theory for 2-D continuous functions is not yet available.

Uniqueness Condition Due to Eisenstein's Theorem

Although most 2-D sampled objects are, as discussed above, uniquely related to the modulus of their Fourier transforms, it is of interest to know conditions that ensure uniqueness. Such a condition was recently put forward by Fiddy *et al.*[13] They considered the class of sampled objects whose support is contained in the union of a rectangle and an isolated point (A) below and to the right of the rectangle, as shown in Fig. 1(a). By way of example, the rectangular region in Fig. 1(a) contains five columns and four rows of points. The object must also be nonzero both at point A and at point B in the lower left corner of the rectangle. Points A and B are referred to as the reference points, and they do not satisfy the holography condition. If these conditions are satisfied, then the Fourier transform of the object satisfies Eisenstein's theorem, making it an irreducible 2-D polynomial and guaranteeing that the solution to the phase retrieval problem is unique. They demonstrated the power of these conditions by reconstruction experiments using the input–output iterative Fourier-transform algorithm.[6,7] First, they performed a reconstruction experiment on the Fourier modulus of a particular object that did not have a reference point A. After 250 iterations, a poor reconstruction resulted. But when a new object was formed by adding a reference point A off its corner making it satisfy the conditions, then a good reconstruction was obtained after only 20 iterations.[13] Note that this does not prove that the original object (without the point A) was nonunique: the failure of the iterative reconstruction algorithm may only be an indication of local minima in the error function. In fact, when the reference point A had a small value, a poor reconstruction was obtained in spite of the fact that irreducibility (and uniqueness) was ensured. Only when a large value for A was used did the reconstruction become easier.[13] Apparently the use of a large enough value for A also ensures that there are no local minima.

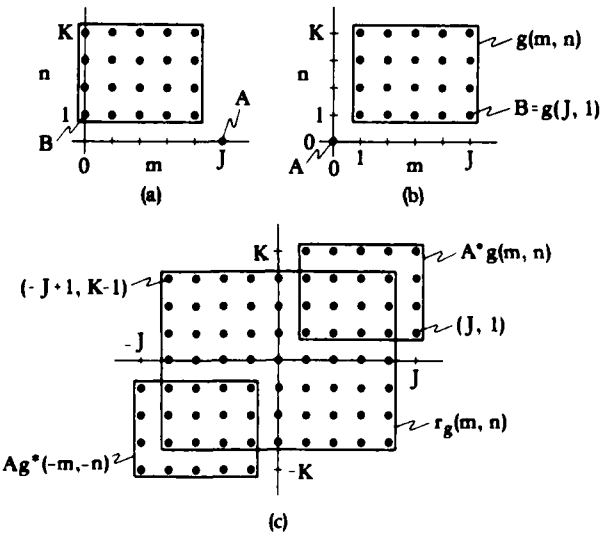


Fig. 1. Fiddy Brames Dainty[13] object. (a) FBD object support having two reference points, A and B; (b) object support assumed; (c) autocorrelation support. The object is uniquely reconstructed from its autocorrelation function.

3. NEW RECONSTRUCTION ALGORITHM

For certain classes of sampled objects having reference points not satisfying the holography condition, we present a new reconstruction algorithm having a fixed number of steps. This new algorithm is related to the Dallas[5] recursive algorithm for phase retrieval from two intensity measurements but requiring only a single intensity measurement (the Fourier modulus) and solving the equations in a certain order such that the problem of a growing tree of solutions[5] is avoided. First the algorithm will be described for the type of object described above, and later for a wider class of objects.

A. Fiddy-Brames-Dainty Objects

For mathematical simplicity, consider a sampled object whose support is contained in the regions shown in Fig. 1(b). Its uniqueness properties are the same as those of the objects considered in Fig. 1(a) since the supports are mirror images of one another. The object can be expressed as in Eq. (4) with $m_0 = n_0 = 0$:

$$f(m,n) = A\delta(m,n) + g(m,n),$$

where $g(m,n)$ is that part of $f(m,n)$ contained in the rectangular region of support, and $A = f(0,0) \neq 0$. In this case, $g(m,n)$ is zero outside $1 \leq m \leq J$ and $1 \leq n \leq K$; and it is assumed that $f(J,1) = g(J,1) = B \neq 0$, and $g(m,K) \neq 0$ for at least one value of m. We will refer to objects satisfying these constraints as Fiddy–Brames–Dainty (FBD) objects having FBD regions of support.

The autocorrelation, $r_f(m,n)$, of $f(m,n)$ is given by the four terms of Eq. (5) with $m_0 = n_0 = 0$, the supports of which are contained in the sets of points illustrated in Fig. 1(c). From this figure, it can be clearly seen that the rightmost column and the uppermost row of $r_f(m,n)$ are simply equal to $A^*g(m,n)$:

$$r_f(J, n) = A^* g(J, n), \qquad n = 1, \ldots, K, \tag{6}$$

$$r_f(m, K) = A^* g(m, K), \qquad m = 1, \ldots, J. \tag{7}$$

Therefore, for $m = J$ and for $n = K$, one can reconstruct $g(m, n)$ to within a constant factor A^* by simple inspection of $r_f(m, n)$. In effect, the holography condition is in force for the row and column opposite reference point A, and that row and that column are reconstructed by using reference point A.

The value of A can be obtained as follows: From Eq. (2), it is seen that there is only one nonzero term in the summation for the upper left corner point in the autocorrelation:

$$r_f(-J + 1, K - 1) = g(1, K)g^*(J, 1) = B^* g(1, K). \tag{8}$$

Also, from Eqs. (6) and (7),

$$r_f(J, 1) = A^* g(J, 1) = A^* B, \tag{9}$$

$$r_f(1, K) = A^* g(1, K). \tag{10}$$

Combining Eqs. (8)–(10) yields (assuming that $r_f(-J + 1, K - 1) \neq 0$,

$$|A|^2 = \frac{r_f(J, 1) r_f^*(1, K)}{r_f^*(-J + 1, K - 1)}. \tag{11}$$

Since without loss of generality we can arbitrarily fix the phase of any one point in $f(m, n)$, we set the phase of A equal to zero; A is then given unambiguously by the positive square root of Eq. (11). If $r_f(-J + 1, K - 1) = 0$, then one can obtain a similar expression for $|A|^2$ using the first nonzero point, $r_f(m, K - 1)$, to the right of $r_f(-J + 1, K - 1)$. Since A is known, $g(J, n)$ and $g(m, K)$ can be determined unambiguously from Eqs. (6) and (7). Note that $B = g(J, 1) = r_f(J, 1)/A^*$.

Having the values of the top row and rightmost column of $g(m, n)$, one can then solve for the leftmost column in the second step of the algorithm. From Eq. (2), the point of the autocorrelation just below $r_f(-J + 1, K - 1)$ has only two nonzero terms,

$$r_f(-J + 1, K - 2) = g(1, K)g^*(J, 2) + g(1, K - 1)g^*(J, 1). \tag{12}$$

Solving,

$$g(1, K - 1) = [r_f(-J + 1, K - 2) - g(1, K)g^*(J, 2)]/B^*, \tag{13}$$

where $g(J, 1) = B$. Since all the quantities of the right-hand side of Eq. (13) are known and $B \neq 0$, one can unambiguously compute $g(1, K - 1)$. Similarly, the next lower point in the autocorrelation is given by

$$r_f(-J + 1, K - 3) = g(1, K)g^*(J, 3) + g(1, K - 1)g^*(J, 2) + g(1, K - 2)g^*(J, 1). \tag{14}$$

Since all the quantities in this linear equation are known except for $g(1, K - 2)$, and since $g(J, 1) \neq 0$, one can solve unambiguously for $g(1, K - 2)$. In a similar fashion, one can recursively solve for all the values $g(1, n)$ (the first column on the left) using the values of $r_f(-J + 1, n - 1)$ in this second step of the reconstruction. In a sense the column $m = 1$ was solved using the latent reference point B, which required the solution of column $m = J$ before it could become effective.

Having the first column on the left and the first column on the right of $g(m, n)$, one can then solve for the second column on the right in the third step, using A as the latent reference point. From Eq. (2), the points of the autocorrelation in column $(J - 1)$ are given by

$$r_f(J - 1, n) = g(J - 1, n)A^* + \sum_{k=n+1}^{K} g(J, k)g^*(1, k - n), \tag{15}$$

for $n = 1, \ldots, K - 1$. Since, for any n, $g(J - 1, n)$ is the only unknown in Eq. (15), and since $A \neq 0$, $g(J - 1, n)$ is uniquely determined from Eq. (15). Thus the values of $g(m, n)$ in column $(J - 1)$ are reconstructed using the values in column $(J - 1)$ of the autocorrelation.

The reconstruction algorithm continues in the manner described above. In the fourth step, one can recursively solve for $g(2, n)$ using the latent reference point B and the values of $r_f(-J + 2, n - 1)$, $n = K - 1, K - 2, \ldots, 2, 1$. In the fifth step, one can solve for $g(J - 2, n)$ using the latent reference point A and the values of $r_f(J - 2, n)$, $n = 1, \ldots, K - 1$. One continues the procedure until all the columns of $g(m, n)$ are reconstructed, giving a complete and unambiguous reconstruction of $g(m, n)$, and therefore of $f(m, n)$.

If $g(1, K) \neq 0$, then one can alternatively use that point as B and perform the reconstruction as described above, but reversing the roles of the rows and columns.

It was recently noted that Eisenstein's theorem allows for the rectangular region of support (see Fig. 1) to be extended over (in the same column as) point A. However, in that case, there is no simple recursive algorithm for reconstructing the object.

B. Support Uniqueness for Fiddy-Brames-Dainty Objects

In the reconstruction method described above, it was implicitly assumed that the support of the object function was known. However, as will be shown by what follows, such an assumption is not necessary, since an FBD object can be shown to be an FBD object from its autocorrelation. In order to use theorems[10] relating to reconstructing the support of an object from the support of its autocorrelation function, during the discussion of the object and autocorrelation supports we assume that the object function is real and nonnegative. (It might happen that what follows may, with appropriate modifications, also be true for complex-valued objects; but this would require further development.)

Given only the support of the autocorrelation, one can usually only put an upper bound on the support of the object.[10] Such upper bounds, sets that can contain translates of the supports of all possible solutions, we refer to as locator sets. One such locator set is the intersection of the autocorrelation support with a translate of itself, where the translate is such that the center of the second autocorrelation support is within the first autocorrelation support.[10] Assuming that $r_f(-J + 1, K - 1) \neq 0$, and translating the one autocorrelation support so that it is centered at $(-J + 1, K - 1)$, one arrives at the locator set shown in Fig. 2 for the case of the FBD object support shown in Fig. 1(b). In addition, since the autocorrelation is $2J + 1$ pixels wide and $2K + 1$ pixels high, the object must be $J + 1$ pixels wide and $K + 1$ pixels high. Since the object support must be contained within the locator set shown in Fig. 2, which is $J + 2$ pixels wide and $K + 2$ pixels high, the object support must include either the lower left point or the upper right point but not both. Keeping either one of these

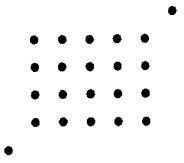


Fig. 2. Locator set containing all possible solutions, used to show that the support solution is unique.

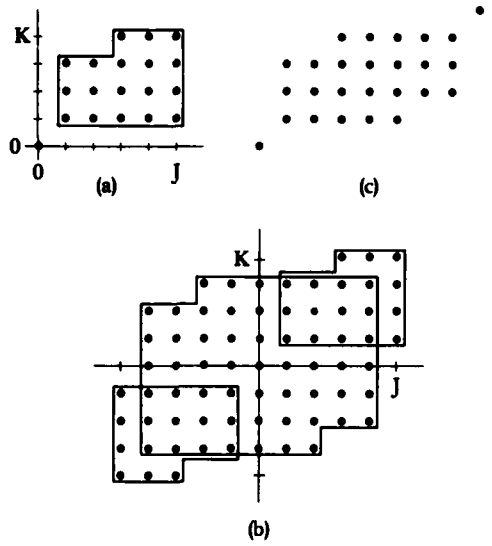


Fig. 3. Alternative case. (a) Object support; (b) autocorrelation support; (c) locator set.

two points and discarding the other, one is left with the support of the object (or the 180° rotated version—the twofold ambiguity). Suppose, on the other hand, that $r_f(-J+1, K-1) = 0$. For example, suppose that the object support is that shown in Fig. 3(a). Then the autocorrelation support is that shown in Fig. 3(b). A locator set, formed by taking the intersection of this autocorrelation support with one translated to be centered at the first nonzero point in row $(K-1)$, is shown in Fig. 3(c). As in the case of Figs. 1 and 2, since the autocorrelation is $2K+1$ pixels high, the object must be $K+1$ pixels high, and therefore either the lower right or the upper left point (but not both) in Fig. 3(c) must be within the object support. Suppose we take the lower left point as being within the object (choosing the upper right point will result in the 180° rotated solution). Then, since the autocorrelation is $2J+1$ pixels wide and therefore the object must be $J+1$ pixels wide, the object must be contained within the first $J+1$ columns on the left of Fig. 3(c), which is just the support of the object as shown in Fig. 3(a). From these examples it can be seen that, in general, if the object is an FBD object, then its support can be reconstructed from the autocorrelation function, from which it is also evident that the object has an FBD support.

C. Triangular Objects

Other types of objects, in addition to FBD objects, can be reconstructed by the recursive method. In this and the next section the reconstruction of two other classes of objects are shown. Consider, for example, objects whose support is contained in the triangular shape shown in Fig. 4(a). Assuming that the object's support is known *a priori*, it has been shown that for this particular object shape the boundaries can be reconstructed in a simple way,[14] assuming $A, B, C \neq 0$. Since the vector spacings between points A and B, B and C, and C and A are each unique, from the corner points in the autocorrelation, as shown in Fig. 4(b), we have

$$r(0, K) = f(0, K)f^*(0, 0) = CA^*, \tag{16a}$$

$$r(J, -K) = f(J, 0)f^*(0, K) = BC^*, \tag{16b}$$

$$r(J, 0) = f(J, 0)f^*(0, 0) = BA^*. \tag{16c}$$

Combining these gives

$$|A|^2 = \frac{r^*(0, K)r(J, 0)}{r(J, -K)}. \tag{17}$$

Without loss of generality the phase of A can be chosen to be zero, and then A is given by the positive square root of Eq. (17). Then we can also compute

$$B = r(J, 0)/A^*, \tag{18a}$$

$$C = r(0, K)/A^*. \tag{18b}$$

Then the values of the leftmost column of the object are given by

$$f(0, n) = r(-J, n)/B^*, \tag{19}$$

the values of the bottom row are given by

$$f(m, 0) = r(m, -K)/C^*, \tag{20}$$

and the values of the diagonal are given by

$$f(m, K-m) = r(m, K-m)/A^*. \tag{21}$$

From this point one could determine the remainder of the object by solving systems of equations,[5,14] but an easier way

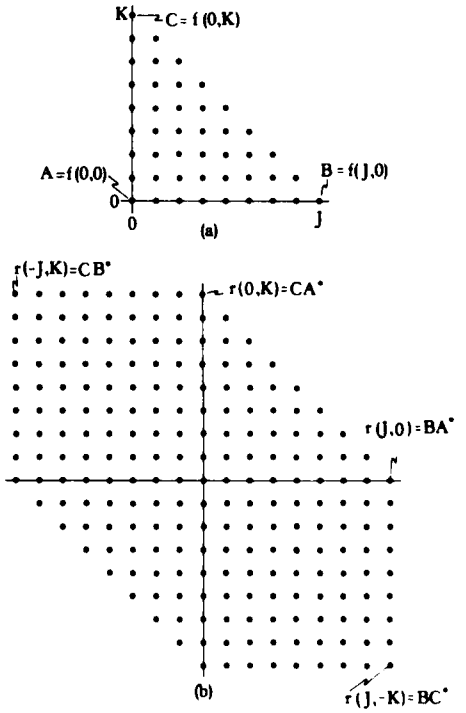


Fig. 4. Triangular-shaped object. (a) Object support; (b) autocorrelation support. The object is uniquely (among triangular-shaped solutions) reconstructed from its autocorrelation function.

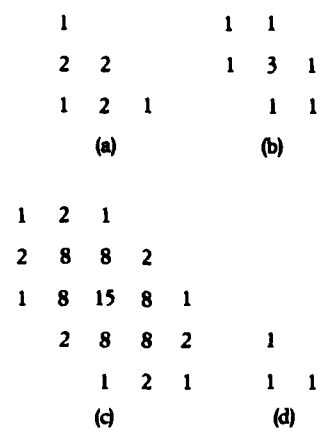


Fig. 5. Specific triangular-shaped object. (a) The object; (b) a second nontriangular-shaped solution; (c) the common autocorrelation function; (d) the function used to synthesize objects shown in (a) and (b).

is possible if one cleverly chooses the order in which the equations are solved. In particular, only one linear equation with one unknown at a time need be solved, and the solution at each step is unique, if one solves in the following order. In a similar manner as was done for the FBD objects, solve for the points in column $m = 1$ using B as a latent reference point, and solve for the points in row $n = 1$ using C as a latent reference point. Next solve for the points in column $m = 2$ using B as a latent reference point, and solve for the points in row $n = 2$ using C as a latent reference point. This procedure is continued until all of $f(m, n)$ is reconstructed. Other orderings for the recursive solution of the equations are also possible.

The solution given above for the triangular-shaped object is unique among objects having that support but may not be unique among all objects. Momentarily restricting $f(m, n)$ to the case of nonnegative objects, one can use the autocorrelation support tri-intersection reconstruction for convex sets[10] to show that there exists a family of object supports that have autocorrelation supports equal to the one shown in Fig. 4(b). One member of that family is the original object support shown in Fig. 4(a). Another member is an object support resembling the autocorrelation support shown in Fig. 4(b) but only half its size. For these latter members there is no simple recursive reconstruction algorithm as there is for the triangular-shaped object.

Further insights can be obtained by analyzing a simple case. A case for which there are exactly two different solutions (not counting 180°-rotated versions) can be obtained by starting with nonsymmetric functions $h_1(x, y)$ and $h_2(x, y)$ whose Fourier transforms are nonfactorable and generating a first object, which is $h_1(x, y)$ convolved with $h_2(x, y)$, and a second object, which is $h_1(x, y)$ convolved with $h_2(-x, -y)$ (i.e., the cross correlation).[15] Two such objects, their common autocorrelation function, and the $h_1(x, y) = h_2(x, y)$ used to generate them are shown in Figs. 5(a) through 5(d), respectively. In this case one obtains the "unique" solution shown in Fig. 5(a) if triangular support is assumed, and the "unique" solution shown in Fig. 5(b) if the only other possible support is assumed.

Since relatively few 2-D objects have factorable Fourier transforms, the ambiguous example shown in Fig. 5 is unusual. If one started with a random object having the same support as the object in Fig. 5(b), and if one incorrectly assumed that the object had the same triangular support as the object in Fig. 5(a), then one would obtain what at first glance would appear to be a triangular-shaped solution. In the process of calculating the solution one would use only the points on the perimeter of the autocorrelation function, with which the "solution" would be consistent. However, on further inspection one would usually find that the triangular-shaped solution is inconsistent with the interior points of the autocorrelation function. Only in the unlikely event that the original object's Fourier transform is factorable would the triangular-shaped solution be completely consistent with the autocorrelation function. Therefore if the given autocorrelation function admits to a possible solution by the recursive method, then one should reconstruct the solution with the assumed support, then compute its autocorrelation function and compare it with the given autocorrelation function to determine whether the assumed support is valid.

D. Another Case

For a final example, consider objects contained within the support shown in Fig. 6(a). Comparing it with Fig. 1(b), it would be a FBD object if it were not for the fact that $B = 0$. Assuming that the support of the object is known, it can be reconstructed by the following recursive steps if points A and $B' \neq 0$ and if either point C' or $C'' \neq 0$. First $f(J, 2), \ldots, f(J, K)$ and $f(2, K), \ldots, f(J-1, K)$ are solved using A as the reference point. A can be determined from an equation similar to Eqs. (11) and (17). Next $C' = f(1, K-1)$, then $f(1, K-2)$, $\ldots$, then $f(1, 2)$ are solved using B' as the latent reference point. Next $f(J-1, 1)$ is solved using C' or C'' as the latent reference point. Next $f(1, 1)$ is solved using B' as the latent reference point. Next $f(J-1, 2), \ldots, f(J-1, K-1)$ are solved using A as the latent reference point. Then the pattern repeats: solve for $f(2, K-1), \ldots, f(2, 2)$ recursively using B', then solve for $f(J-2, 1)$ using C' or C'', then solve for $f(2, 1)$

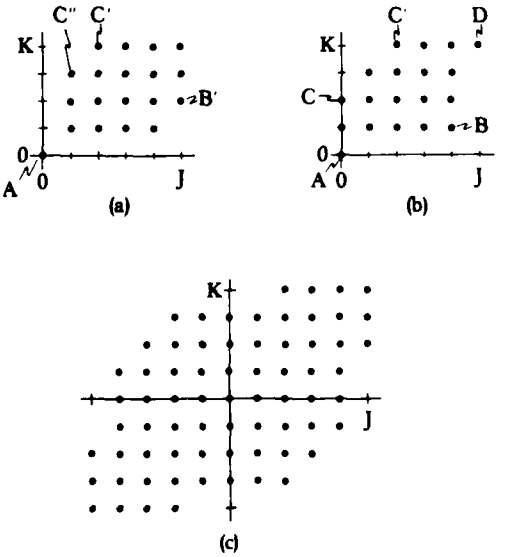


Fig. 6. Another case related to FBD objects. (a) Object support; (b) alternative support reconstruction; (c) autocorrelation support. The object is reconstructed from its autocorrelation function, with two solutions.

using B', then solve for $f(J-2,2),\ldots,f(J-2,K-1)$ using A, etc., until all the columns are solved.

The solution for this object is unique among objects having support contained within the support shown in Fig. 6(a). However, another support may also be possible. In a manner similar to that used in connection with Figs. 1–3, the possible support solutions can be narrowed down to those of Fig. 6(a) and Fig. 6(b), given the autocorrelation support shown in Fig. 6(c). For the support shown in Fig. 6(b) one can reconstruct the object unambiguously by solving a proper sequence of equations using latent reference points A, B, C, C', and D. Therefore, given the autocorrelation function whose support is shown in Fig. 6(c), at most two (and more probably only one) solutions are possible, and each can be reconstructed using a simple recursive algorithm depending on the support shown in either Fig. 6(a) or 6(b).

4. CONCLUSIONS

A simple recursive algorithm has been devised for reconstructing an object from its autocorrelation function (or its Fourier modulus). It works for several types of sampled objects having latent reference points, including those satisfying the conditions described by FBD. The manner in which the algorithm results in a unique solution constitutes a proof of uniqueness for FBD objects (but not necessarily for all objects whose Fourier transforms satisfy Eisenstein's theorem). One of the principal lessons learned here is that the detailed shape of the boundary of an object plays a crucial role in determining the uniqueness of the solution to the phase-retrieval problem.

One might be able to use this method for continuous objects (as opposed to inherently sampled objects) if a dense enough sampling of the autocorrelation is available.[5]

Since the algorithm involves repeatedly taking differences and dividing by the values of the latent reference points, it may be sensitive to noise and may require latent reference points having large values for an accurate reconstruction. (This may be related to the fact that a large value of A was required for a successful reconstruction using the iterative Fourier-transform algorithm.[13]) Not all the (nonsymmetric) points in the autocorrelation are used by this algorithm; improved accuracy should be expected if the reconstruction algorithm were modified to use also those additional points. Those additional points may also be used to distinguish whether assumptions about the support of the object (when more than one support solution is possible) are valid. For the best results one should finish the reconstruction by using the output of this reconstruction method as the initial input to the iterative Fourier-transform algorithm,[6,7] which finds a solution that is most consistent with both the measured data and the *a priori* constraints.

The reconstruction algorithm proposed here is applicable to only a relatively small number of types of objects. However, the approach of carefully selecting the order in which the equations are solved should be helpful in the more general use of Dallas's method by limiting the growth of the tree of solutions.[5]

ACKNOWLEDGMENT

Helpful discussions with T. R. Crimmins are gratefully acknowledged. This research was supported by the U.S. Air Force Office of Scientific Research under contract F49620-82-K-0018.

REFERENCES

1. R. N. Bracewell, *The Fourier Transform and Its Applications*, 2nd ed. (McGraw-Hill, New York, 1978).
2. C. Y. C. Liu and A. W. Lohmann, "High resolution image formation through the turbulent atmosphere," Opt. Commun. **8**, 372–377 (1973).
3. J. W. Goodman, "Analogy between holography and interferometric image formation," J. Opt. Soc. Am. **60**, 506–509 (1970).
4. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972); W. O. Saxton, *Computer Techniques for Image Processing in Electron Microscopy* (Academic, New York, 1978); R. H. Boucher, "Convergence of algorithms for phase retrieval from two intensity distributions," Proc. Soc. Photo-Opt. Instrum. Eng. **231**, 130–141 (1980).
5. W. J. Dallas, "Digital computation of image complex amplitude from image- and diffraction-intensity: an alternative to holography," Optik **44**, 45–59 (1975).
6. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
7. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. **3**, 27–29 (1978); J. R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. **18**, 529–534 (1979).
8. P. J. Napier and R. H. T. Bates, "Inferring phase information from modulus information in two-dimensional aperture synthesis," Astron. Astrophys. Suppl. **15**, 427–430 (1974).
9. G. H. Stout and L. H. Jensen, *X-Ray Structure Determination* (Macmillan, London, 1968).
10. J. R. Fienup, T. R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation," J. Opt. Soc. Am. **72**, 610–624 (1982).
11. Yu. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. **30**, 304–308 (1979).
12. A. Walther, "The question of phase retrieval in optics," Opt. Acta **10**, 41–49 (1963); E. M. Hofstetter, "Construction of time-limited functions with specified autocorrelation functions," IEEE Trans. Inf. Theory **IT-10**, 119–126 (1964).
13. M. A. Fiddy, B. J. Brames and J. C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. **8**, 96–98 (1983).
14. M. H. Hayes and T. F. Quatieri, "The importance of boundary conditions in the phase retrieval problem," IEEE Trans. Acoust. Speech Signal Process. **ASSP-82**, 1545 (1982).
15. J. R. Fienup, "Image reconstruction for stellar interferometry," in *Current Trends in Optics*, F. T. Arecchi and F. R. Aussenegg, eds. (Taylor and Francis, London, 1981), pp. 95–102.

Appendix C

HOLOGRAPHIC RECONSTRUCTION WITH LATENT REFERENCE POINTS

J.R. Fienup

Presented at the Annual Meeting of the Optical Society of America, New Orleans, Louisiana, October 1983; Abstract: Journal of the Optical Society of America 73, 1861 (December 1983).

MH9. Holographic Reconstruction with Latent Reference Points.* J. R. FIENUP, *Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107* — In the image plane of a Fourier-transform hologram, one finds the autocorrelation of the object-plane distribution (the object plus a reference point), which includes the autocorrelation of the object, the crosscorrelation of the object with the reference point (i.e., the desired image), and the conjugate image. When the reference point is insufficiently offset from the object, then a straightforward reconstruction is frustrated by the overlap of the desired image with the autocorrelation term. One then must solve the Fourier-phase retrieval problem or equivalently reconstruct the object-plane distribution from its autocorrelation function. For certain cases there is a unique relationship between the Fourier-plane intensity and the object-plane distribution, even when the reference point is close to the object (hence it is only a latent reference point). Examples of this are object-plane distributions whose Fourier transforms satisfy Eisenstein's criterion.[1] For these and certain other types of object-plane distributions, one can digitally reconstruct the object-plane distribution from its autocorrelation function using a recursive algorithm. This also constitutes a proof of the uniqueness of phase retrieval for these types of object-plane distributions. The algorithm is similar to Dallas' algorithm except that it involves solving only one linear equation at a time. It has applications in holography, astronomy, and wave-front sensing. (13 min.)

* This research was supported by the U.S. Air Force Office of Scientific Research

[1] M. A. Fiddy, B. J. Brames, and J. C. Dainty, Opt. Lett. 8, 96 (1983).

Appendix D

EXPERIMENTAL EVIDENCE OF THE UNIQUENESS OF PHASE RETRIEVAL FROM INTENSITY DATA

J.R. Fienup

Published in Indirect Imaging, Proceedings of the URSI/IAU Symposium, 30 August to 2 September 1983, Sydney, Australia, ed. J.A. Roberts (Cambridge University Press, 1984), pp. 99-109.

EXPERIMENTAL EVIDENCE OF THE UNIQUENESS OF PHASE RETRIEVAL FROM INTENSITY DATA

J.R. Fienup
Environmental Research Institute of Michigan
P.O. Box 8618, Ann Arbor, Michigan 48107, USA

Summary. An increasing body of theory indicates that the phase retrieval problem usually has a unique solution for 2-D objects. In this paper experimental reconstruction results that support the uniqueness theory are shown.

1 INTRODUCTION

In both optical and radio astronomy, sometimes one can accurately obtain the modulus of the Fourier transform (i.e., the magnitude of the complex visibility function) of an image, but not the Fourier phase. In order to obtain an image it then becomes necessary to retrieve the Fourier phase. Since the autocorrelation function can be computed as the inverse Fourier transform of the squared Fourier modulus, the problem is equivalent to reconstructing an image from its autocorrelation.

In this paper we are concerned with the phase retrieval problem in optical astronomy, in which case one cannot rely on such aids as closure phase (Jennison 1958). However the results shown here do have relevance to radio astronomy as well.

Several methods have been put forward for solving the phase retrieval problem (Liu & Lohmann 1973; Napier & Bates 1974; Frieden & Currie 1976; Baldwin & Warner 1978; Fienup 1978, 1979, 1982; Bates et al. 1982a). In addition there are a number of reconstruction techniques that depend on the specific method of data collection, for example, astronomical speckle interferometry (Bates 1982b). Of the methods that would work for the most general case, the iterative input-output Fourier transform algorithm (Fienup 1978, 1979, 1982) appears to be the most practical.

When any of the reconstruction algorithms finds a solution, the question remains: is it the only (unique) solution or is it one of many possible (ambiguous) solutions? In Section 2 the theory of the uniqueness will be briefly reviewed. Then in Section 3 experimental reconstruction results will be shown that are consistent with the theory that the 2-D case is usually unique. In addition, experimental reconstruction results are shown that indicate that noise in the Fourier modulus data does not radically change the uniqueness of the solution.

2 UNIQUENESS THEORY

When we speak of the reconstruction being unique or ambiguous, we ignore translations and 180° rotations since neither of these operations affects the Fourier modulus. Here we are also assuming that the object has a finite spatial (or angular) extent.

The one-dimensional (1-D) phase retrieval problem has long been known to be highly ambiguous (Walther 1963). Only for the special cases of objects known to consist of sufficiently separated parts or nonnegative objects having sufficiently separated parts is the 1-D phase retrieval problem usually unique (Greenaway 1977; Crimmins & Fienup 1983).

The 2-D case is quite different. This can best be understood from the theory developed by Bruck and Sodin (1979). They considered the special case of an object sampled on a rectangular lattice. For the 1-D case the Fourier transform can then be expressed as a polynomial of order M of a single complex variable, and such a polynomial can always be factored into M irreducible factors (by the fundamental theorem of algebra). They showed that this implies that in the 1-D case there are 2^{M-1} possible solutions, although not all of those solutions would satisfy a nonnegativity constraint (Bates 1969). On the other hand, polynomials of two complex variables having arbitrary coefficients are only rarely factorable. Consequently the 2-D case is usually unique. Although the 2-D theory for continuous functions has not yet been fully developed, it is likely that a similar result will hold.

Of course one can always fabricate 2-D examples that are not unique. An example is an object formed by convolving two nonnegative functions. A second object, formed by convolving the first nonnegative function with an inverted (i.e., rotated by 180°) version of the second nonnegative function, has the same Fourier modulus as the first object. Another method of synthesizing ambiguous cases was given by Huiser and van Toorn (1980). However, these fabricated ambiguous objects are very special cases--most 2-D objects do not fit into these categories.

There are also a number of classes of objects for which the phase retrieval problem is known to be unique (as opposed to just being usually unique). For example, if the object includes an unresolved (delta-function-like) point far enough away from the rest of the object, then the autocorrelation includes the rest of the object as one of its terms (Liu & Lohmann 1973), analogous to holography. It has also been recently discovered that for objects having a special support there is a unique reconstruction even if the reference points are very close to the rest of the object (Fiddy et al. 1983). The support of an object is the set of points over which it is nonzero, i.e., its shape. Also using latent reference points it can be shown that these and other objects having certain supports can be uniquely reconstructed from their Fourier modulus (Fienup 1983a). These recent results point to the importance of the support of an object in determining whether the object can be uniquely reconstructed from its Fourier modulus. Methods for recon-

structing support information without resorting to a complete reconstruction are also being investigated (Fienup et al. 1982).

3 UNIQUENESS EXPERIMENTS

3.1 Iterative reconstruction algorithm

One approach to determining whether most objects of interest are uniquely reconstructable from their Fourier modulus is to perform a number of reconstruction experiments. This is now possible due to the existence of a practical reconstruction algorithm, the iterative Fourier transform algorithm (Fienup 1978, 1979, 1982).

The iterative Fourier transform algorithm uses all the available measurements and a priori information to arrive at a solution. In the Fourier domain one has the measured Fourier modulus data, which is an estimate of the true modulus of the Fourier transform of the object. In the object domain one has the a priori constraint that the object's spatial (or angular) brightness distribution is a nonnegative function. From the Fourier modulus data one can compute an estimate of the object's autocorrelation function. From the autocorrelation one can place upper bounds on the diameter of the object (only in special cases can the support of the object be readily determined from the support of its autocorrelation) (Fienup et al. 1982).

The iterative Fourier transform algorithm is a modification of the Gerchberg-Saxton (1972) algorithm that has been used in electron microscopy and for other applications (Fienup 1983b). The simplest version of the iterative algorithm consists of the four following steps. (1) An estimate of the object (an input image) is Fourier transformed. (2) The resulting Fourier-domain function is forced to conform to the measurements by replacing the computed Fourier modulus with the measured Fourier modulus. (3) The result is inverse Fourier transformed, yielding an output image. (4) A new input image is formed by forcing the output image to conform to the object-domain constraints, i.e., it is set equal to zero where it is negative or where is exceeds the known diameter (i.e., the support constraint). This algorithm, which we call the error-reduction algorithm, can be proven to converge in the sense that the error at the k^{th} iteration is always less than or equal to the error at the $(k - 1)^{th}$ iteration. Here the error is defined as the amount by which the computed Fourier modulus differs from the measured Fourier modulus or as the amount by which the output image violates the object-domain constraints. However, in practice the error-reduction algorithm usually converges so slowly that it is impractical for this application (Fienup 1982).

Fortunately there exist a number of accelerated versions of the algorithm which converge in a reasonable number of iterations. To date the fastest version of the algorithm is the hybrid input-output algorithm. Its first three steps are identical to those of the error-reduction algorithm described above. The fourth step of the hybrid input-output

algorithm consists of forming a new input image that is equal to the output image wherever the output image satisfies the constraints, and is equal to the previous input image minus a constant factor times the output image wherever the output image violates the constraints. Any value between 0.5 and 1.0 works well for constant factor, which is similar to a negative feedback parameter.

In one series of trials, the algorithm was run on a fabricated Fourier modulus which was known to have two solutions. One of the two solutions was reconstructed in about half of the trials and the other solution was reconstructed in the other half of the trials. Which of the two solutions was obtained depended on the array of random numbers used as the initial input to the algorithm. Therefore we believe that if there are multiple solutions, then the algorithm is equally likely to find any one of them (if the initial input is sufficiently random and unbiased), and if run enough times with different initial inputs, it will probably find all of them. In a practical reconstruction situation in which the solution is not known beforehand, if one were to run the algorithm two or three times, each time using a different array of random numbers for the initial input, and if the reconstructed images were the same each time, then one would be highly confident that one had found the solution and that it is unique (Fienup 1979).

A problem with experimental reconstruction experiments is that there is no guarantee that the iterative algorithm will converge to <u>any</u> solution, even when an accelerated version of the algorithm is used. One can think of the reconstruction algorithm as an iterative search through an N^2-dimensional parameter space (each dimension or parameter corresponding to the value of one of the pixels of the image), seeking to minimize the error of the estimate. While searching for the global minimum of the error, the algorithm could stagnate at a local minimum of the error in that N^2-dimensional space. The likelihood of stagnation and the success of the algorithm depend on the N^2-dimensional topography of the error function, which varies from one type of object to another. Therefore, for particularly difficult objects, i.e., ones for which the error has many local minima, one may not be able to test for uniqueness since the reconstruction algorithm fails. Fortunately such a problem has occurred only occasionally for the types of objects examined.

One particular convergence problem has occurred on several occasions. Sometimes the algorithm stagnates at a deep local minimum at which the output image resembles the original object but with a pattern of stripes superimposed. A similar phenomenon has occurred in other reconstruction situations (Cornwell 1983). In most cases the stripes are of low contrast, superimposed on an otherwise excellent reconstructed image, and are of little concern. In other cases the stripes are of high enough contrast to be objectionable. When the Fourier modulus data is sufficiently noisy, then the stripes do not appear (Feldkamp & Fienup 1980). The nature of the stripes is as yet not fully understood and methods of avoiding them remain to be developed.

An example of the stripes phenomenon is shown in Figure 1. Figure 1(a) shows the original object and Figure 1(b) shows a reconstructed image, which appears to be quite faithful. Figure 1(c) shows the same reconstructed image, but heavily overexposed, in order to emphasize the low-contrast vertical stripes that are present, although difficult to discern, in the image. Figures 1(d-f) show the overexposed reconstructed images resulting from three other trials of the algorithm, each of which was initialized with a different array of random numbers. In each of these three cases the reconstructed image contains a more easily discernable pattern of stripes, but the spatial frequencies and orientations of the stripes are different in each case. The stripes extend throughout image space (although they are weaker away from the support of the object), and therefore by inspection of the reconstructed images it is possible to determine that the stripes are an artifact rather than a true feature of the object. Furthermore, it is possible to discern the true image from the stripes since the stripes change from one reconstruction to the next, but the true features of the object are present in all the reconstructed images.

3.2 Experimental uniqueness results for various objects

The iterative reconstruction algorithm was used to reconstruct a number of different objects from their Fourier modulus. The objects examined are of a very practical and interesting class: digitized photographs of satellites. They also share a feature that we suspect makes them "good" objects to reconstruct: they have interesting (i.e., complicated) shapes.

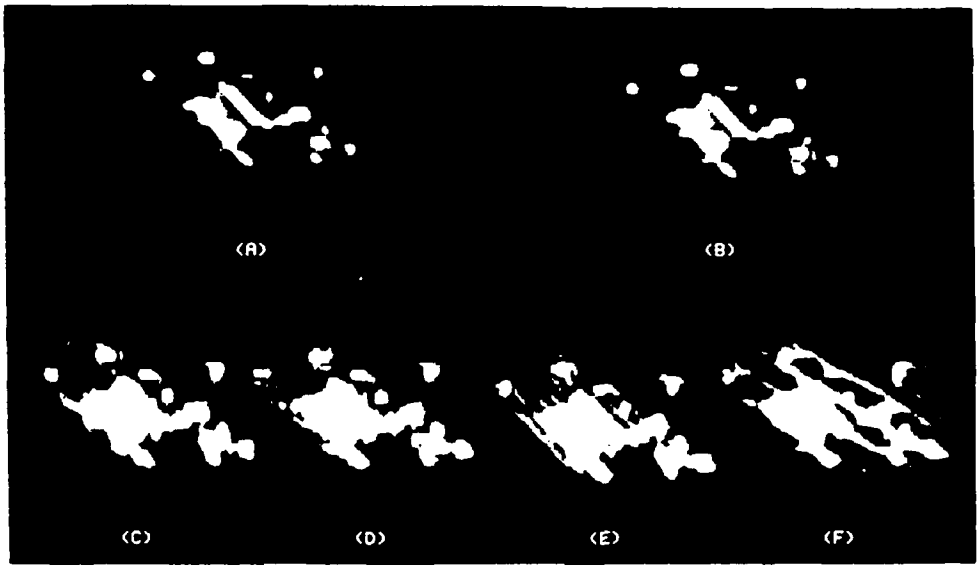
A typical result is shown in Figure 2, in which (a) is the original object and (b) is the reconstructed image (Fienup 1981). For this and almost all of the cases examined, the reconstructed image looks much like the original object except for differences that could be attributed to stripes. For example, horizontal stripes are evident over portions of the reconstructed image shown in Figure 2(b). Therefore, except for the presence of the stripes artifact which we believe is a characteristic of a local minimum rather than an inherent ambiguity, most objects of this type are uniquely related to their Fourier modulus.

There are exceptions, however. Figure 3 shows one case that worked particularly poorly. The object shown in Figure 3(a) is nearly centrosymmetric. Figure 3(b) shows the reconstructed image, which is not very faithful. This particular case has similarities with the ambiguous case fabricated by Huiser and van Toorn (1980). From this we see that, although ambiguous cases may be unusual, they are by no means nonexistent in the real world.

3.3 Experimental uniqueness in the presence of noise

As with any reconstruction method the sensitivity of the algorithm to noise is a major point of concern. Reconstruction results using noisy Fourier modulus data have shown that the iterative Fourier

Figure 1. (a) Original object; (b) image reconstructed from Fourier modulus using iterative algorithm; (c)-(f) four images reconstructed using different starting inputs--these pictures were intentionally overexposed in order to emphasize the stripes.



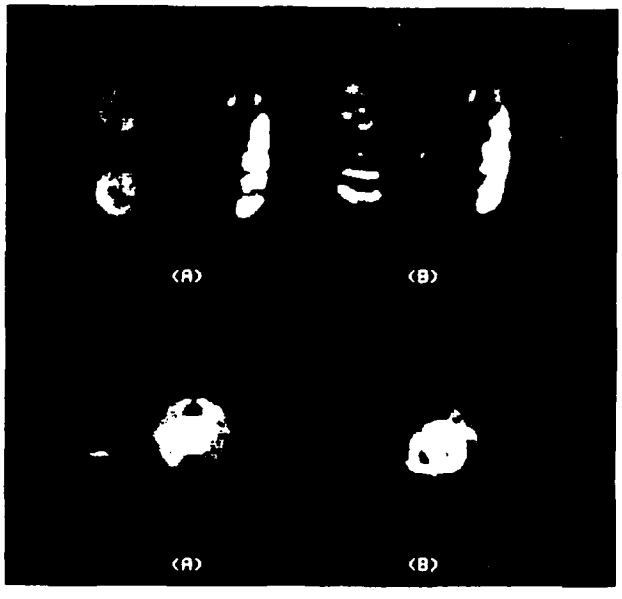


Figure 2. (a) A typical object; (b) image reconstructed from Fourier modulus.

Figure 3. (a) An atypical object, for which the reconstructed image (b) does not resemble the object.

transform algorithm is not highly sensitive to noise (Fienup 1978, 1979). In this section the results of a systematic study of the noise sensitivity of the reconstruction (Feldkamp & Fienup 1980) are summarized.

When noise is present in the Fourier modulus data, then there is generally no solution that is completely consistent with both the measured data and the constraints. For example, an autocorrelation function computed from a noisy Fourier modulus would be very likely to have some negative values for the largest separations. Obviously no nonnegative object can have an autocorrelation having negatives; therefore there could be no nonnegative object consistent with the noisy Fourier modulus. Nevertheless the algorithm searches for a solution that is most consistent with the measured data and constraints, and in doing so it can arrive at a useful image.

Fourier modulus data was simulated to have the type of noise that would be present in astronomical speckle interferometry. The object shown in Figure 1(a) was convolved with 156 different point-spread functions to produce 156 different blurred images. Each of the point-spread functions represents a different realization of the blurring due to the turbulent atmosphere. The widths of the point-spread functions were comparable to the width of the object. The blurred images were then subjected to a Poisson noise process to simulate the effects of photon noise. The degraded images were then processed to produce a noisy Fourier modulus estimate by Labeyrie's (1970) method, as modified by Goodman and Belsher (1976) to eliminate the bias noise term from the squared Fourier modulus.

Figure 4 shows a noise-free Fourier modulus (a) and three examples of the simulated noisy Fourier modulus estimates (b)-(d) with increasing noise. Figure 5 shows the original undegraded object (a) and three images (b)-(d) reconstructed from the respective noisy Fourier modulus estimates of Figure 4. For the case shown in Figures 4(b) and 5(b), which represent a realistic amount of noise for this situation, the normalized rms error of the Fourier modulus estimate was 2.9% and the reconstructed image is very good. For the case shown in Figures 4(c) and 5(c), only 1/50 as many photons were assumed to be available, and the rms error of the Fourier modulus estimate is a very poor 32%; nevertheless the reconstructed image still retains some recognizable features. In the case shown in Figures 4(d) and 5(d), an extreme amount of noise was present, and the rms error of the Fourier modulus estimate is near 100%; since this Fourier modulus estimate does not resemble the true Fourier modulus, then, as one would expect, the reconstructed image does not resemble the original object.

4. CONCLUSIONS

Theory, which points toward the conclusion that a 2-D object of finite extent is ordinarily uniquely related to the modulus of its Fourier transform, has been supported by experimental reconstruction re-

Figure 4. Fourier modulus estimates with noise, having rms error (a) 0%, (b) 2.9%, (c) 32%, (d) ~100%.

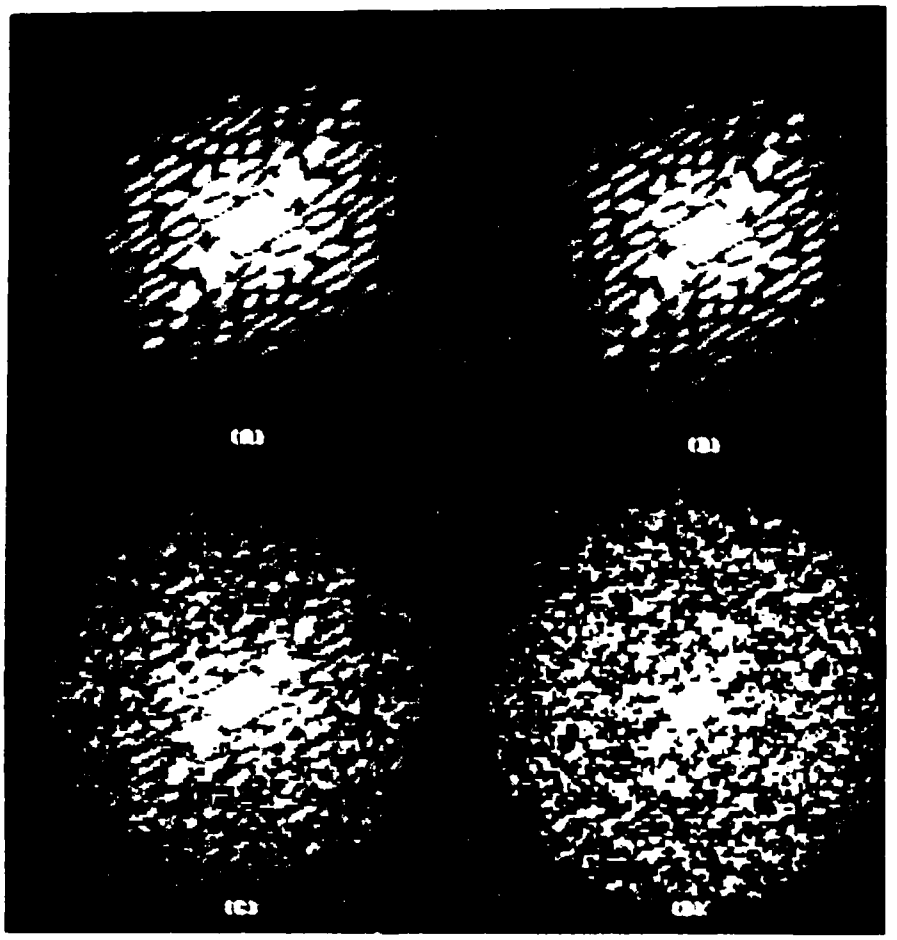
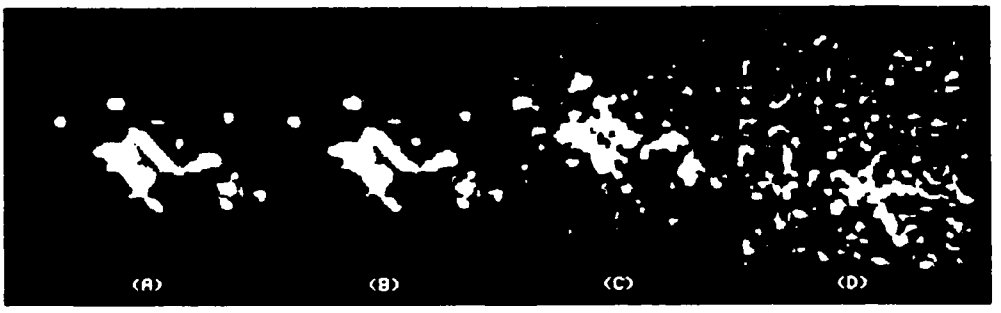


Figure 5. Images reconstructed from noisy Fourier modulus estimates shown in Figure 4.



sults. The vast majority of reconstructed images of satellites resemble the original objects from which the Fourier modulus was computed. Furthermore, contrary to some predictions (Huiser & van Toorn 1980), the uniqueness properties do not change radically when noise is present. Rather, as more noise is introduced into the Fourier modulus estimate, the reconstructed image simply becomes correspondingly noisier, and degrades in a gradual manner.

ACKNOWLEDGEMENT

This research was supported by the U.S. Air Force Office of Scientific Research and Rome Air Development Center.

REFERENCES

Baldwin, J.E. & Warner, P.J. (1978). Phaseless aperture synthesis. Mon. Not. R. Astr. Soc. 182, 411-422.

Bates, R.H.T. (1969). Contributions to the theory of intensity interferometry. Mon. Not. R. Astr. Soc. 142, 413-428.

Bates, R.H.T., et al. (1982a). Fourier phase problems are uniquely solvable in more than one dimension. I: Underlying theory. Optik 61, 247-262; ... II: One-dimensional considerations. Optik 62, 131-142; ... III: Computational examples for two dimensions. Optik 62, 219-230.

Bates, R.H.T. (1982b). Astronomical speckle imaging. Physics Reports (Reviews Sect. of Phys. Lett.) 90, 203-297.

Bruck, Yu.M. & Sodin, L.G. (1979). On the ambiguity of the image reconstruction problem. Opt. Commun. 30, 304-308.

Cornwell, T.J. (1983). A method of stabilizing the CLEAN algorithm. submitted to Astron. Astrophys.

Crimmins, T.R. & Fienup, J.R. (1983). Uniqueness of phase retrieval for functions with sufficiently disconnected support. J. Opt. Soc. Am. 73, 218-221.

Feldkamp, G.B. & Fienup, J.R. (1980). Noise properties of images reconstructed from Fourier modulus. In 1980 International Optical Computing Conference. Proc. SPIE 231, 84-93.

Fiddy, M.A., Brames, B.J., & Dainty, J.C. (1983). Enforcing irreducibility for phase retrieval in two dimensions. Opt. Lett. 8, 96-98.

Fienup, J.R. (1978). Reconstruction of an object from the modulus of its Fourier transform. Opt. Lett. 3, 27-29.

Fienup, J.R. (1979). Space object imaging through the turbulent atmosphere. Opt. Eng. 18, 529-534.

Fienup, J.R. (1981). Fourier modulus image construction. Rept. RADC-TR-81-63.

Fienup, J.R. (1982). Phase retrieval algorithms: a comparison. Appl. Opt. 21, 2758-2769.

Fienup, J.R. (1983a). Reconstruction of objects having latent reference points. To appear in J. Opt. Soc. Am. 73, November.

Fienup, J.R. (1983b). Reconstruction and synthesis applications of an iterative algorithm. In Transformations in Optical Signal Processing, ed. W.T. Rhodes, J.R. Fienup & B.E.A. Saleh. Bellingham, Washington: SPIE.

Fienup, J.R., Crimmins, T.R., & Holsztynski, W. (1982). Reconstruction of the support of an object from the support of its autocorrelation. J. Opt. Soc. Am. 72, 610-624.

Frieden, B.R. & Currie, D.G. (1976). On unfolding the autocorrelation function. J. Opt. Soc. Am. 66, 1111 (Abstract).

Gerchberg, R.W. & Saxton, W.O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik 35, 237-246.

Goodman, J.W. & Belsher, J.F. (1976). Fundamental limitations in linearly invariant restoration of atmospherically degraded images. In Imaging through the atmosphere. Proc. SPIE 75, 141-154.

Greenaway, A.H. (1977). Proposal for phase recovery from a single intensity distribution. Opt. Lett. 1, 10-12.

Huiser, A.M.J. & van Toorn, P. (1980). Ambiguity of the phase-reconstruction problem. Opt. Lett. 5, 499-501.

Jennison, R.G. (1958). Mon. Not. Roy. Astr. Soc. 165, 25.

Labeyrie, A. (1970). Attainment of diffraction limited resolution in large telescopes by Fourier analysing speckle patterns in star images. Astron. and Astrophys. 6, 85-87.

Liu, C.Y.C. & Lohmann, A.W. (1973). High resolution image formation through the turbulent atmosphere. Opt. Commun. 8, 372-377.

Napier, P.J. & Bates, R.H.T. (1974). Inferring phase information from modulus information in two-dimensional aperture synthesis. Astron. Astrophs. Suppl. 15, 427-430.

Walther, A. (1963). The question of phase retrieval in optics. Optica Acta 10, 41-49.

Appendix E

COMMENTS ON "THE RECONSTRUCTION OF A MULTIDIMENSIONAL SEQUENCE FROM THE PHASE OR MAGNITUDE OF ITS FOURIER TRANSFORM"

J.R. Fienup

Reprinted from IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-31, 738-739 (June 1983).

Comments on "The Reconstruction of a Multidimensional Sequence from the Phase or Magnitude of Its Fourier Transform"

J. R. FIENUP

Abstract—**When one imposes a nonnegativity constraint, one usually can reconstruct a two-dimensional sequence of finite support from the modulus of its Fourier transform using an iterative algorithm, even when the initial estimate is an array of random numbers.**

In a recent paper,[1] the description of an iterative algorithm for reconstructing a sequence from the magnitude of its Fourier transform unintentionally gives the appearance of discussing an algorithm published earlier [1]. In the following,

Manuscript received July 13, 1982; revised November 29, 1982. This work was supported in part by the Air Force Office of Scientific Research.

The author is with the Environmental Research Institute of Michigan, Ann Arbor, MI 48107.

[1] M. H. Hayes, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 140-154, Apr. 1982.

0096-3518/83/0600-0738$01.00

the differences between the algorithm and experiments described by Hayes[1] and those published earlier [1] are clarified.

Hayes[1] reviews both the problem of reconstructing a sequence from the phase of its Fourier transform and the problem of reconstructing a sequence from the magnitude of its Fourier transform. For the latter problem, he describes an iterative algorithm for solving the problem as follows. "Specifically, this algorithm involves the repeated Fourier transformation between the time and frequency domains where, in each domain, the known information about the desired sequence is imposed on the current estimate. In the time domain, for example, a sequence is constrained to have a given region of support whereas in the frequency domain, the sequence is constrained to have a given transform magnitude."[1] He then shows examples where the algorithm described above fails. This failure should not reflect poorly on the earlier work [1] since the algorithm described in the quotation above and the experiments performed by Hayes differ in important ways from the earlier work. In Hayes experiments, both the type of information which was assumed to be known and the reconstruction algorithm which was used differed from those of the earlier work [1].

Hayes is correct in stating[1] that the magnitude of the Fourier transform is insufficient to uniquely specify a sequence; additional information or constraints are required. Depending on the application, one often has available additional information or constraints, and a reconstruction may then be possible [2], [3]. Two important constraints which often occur (as in astronomy) are a known support (or bounds on the support) of a sequence, and the constraint that the sequence be nonnegative [4]. Unlike the algorithm used by Hayes,[1] the iterative method described earlier [1] primarily uses the *nonnegativity* constraint. Using the iterative algorithm, we have been very successful in reconstructing two-dimensional nonnegative sequences from their Fourier magnitude [1]-[6]. In this case, the sequences must have finite support, but it is possible to reconstruct them even when the support is not known. Except for special cases, it is not possible to determine the support of a sequence from the support of its autocorrelation (which is the inverse Fourier transform of the squared Fourier magnitude) [7], so the support information is usually not available anyway. One can only place upper bounds on the support [7]. If an upper bound on the support is utilized during the iterations, then the algorithm converges faster (in about 100 or 200 iterations for our work) than when using only the nonnegativity constraint (in which case we found that several hundred iterations are required).

Unlike the algorithm used by Hayes, the iterative algorithm described earlier [1] does not simply satisfy the constraints (nonnegativity and bounds on the support) in the time-domain step of the iteration. Such an algorithm, which we refer to as the error-reduction algorithm, was discussed earlier [1] where it is noted that, "For the present application, the error-reduction approach requires an impractically large number of iterations for convergence." It is only a version of the input-output algorithm [1]-[6] which is capable of converging in 100 or so iterations.

Hayes found that "... if the initial estimate used in the iteration has a Fourier transform with the correct magnitude and either zero phase or random phase, then the iteration will not generally converge to the correct sequence."[1] However, using the input-output algorithm with a nonnegativity constraint, we obtained good reconstruction results when the algorithm was initialized with arrays of random numbers [1]-[6]. The algorithm has also been shown to be surprisingly insensitive to noise [5].

When the error-reduction algorithm was used with a nonnegativity constraint (as well as a support constraint), it took *many thousands* of iterations for convergence [3], [6]. Therefore, if one were to employ the error-reduction algorithm without a nonnegativity constraint, then one would expect convergence to take much longer, if it ever converges. Consequently, it is consistent with our experience that the type of reconstruction experiments performed by Hayes would be unsuccessful.

Of course, there are situations for which the nonnegativity constraint does not apply. Then one might wonder whether it is possible to reconstruct a sequence of finite support from its Fourier magnitude. Theory ([8], Hayes[1]) seems to indicate that the solution will usually be unique. However, as shown by Hayes, the error-reduction algorithm is not a practical approach to finding the solution. One might possibly succeed using an accelerated algorithm, such as the input-output algorithm or a gradient search method [6], but this is an area that needs further work.

It should also be noted that in the phase retrieval problem of X-ray crystallography, one reconstructs the three-dimensional electron density function from its Fourier magnitude. For that problem, one has the constraints that the electron density is nonnegative and that it consists of a discrete number of atoms. For that problem, a number of reconstruction methods have been developed [9]. For the phase retrieval problem in electron microscopy, for which both the wave function and its Fourier transform are complex valued, one has the additional constraint of knowing the magnitude of the wave function. For that problem, the error-reduction algorithm has been shown to perform very well [10], [11].

In conclusion, Hayes' remark that "... even for those sequences which are uniquely defined by their magnitude, it appears that a practical algorithm is yet to be developed for reconstructing a sequence from only its magnitude"[1] is strictly true when no other information is available; however, for a number of important applications, there is auxiliary information, such as a nonnegativity constraint, and practical reconstruction algorithms do exist.

REFERENCES

[1] J. R. Fienup, "Space object imaging through the turbulent atmosphere," *Opt. Eng.*, vol. 18, pp. 529-534, Sept.-Oct. 1979.

[2] ——, "Iterative method applied to image reconstruction and to computer-generated holograms," *Opt. Eng.*, vol. 19, pp. 297-305, May-June 1980.

[3] ——, "Reconstruction and synthesis applications of an iterative algorithm," in *Transformations in Optical Signal Processing*, W. T. Rhodes, J. R. Fienup, and B. E. A. Saleh, Eds. Bellingham, WA: Soc. Photo-Opt. Instrumen. Eng., 1983.

[4] ——, "Reconstruction of an object from the modulus of its Fourier transform," *Opt. Lett.*, vol. 3, pp. 27-29, July 1978.

[5] G. B. Feldkamp and J. R. Fienup, "Noise properties of images reconstructed from Fourier modulus," in *Proc. 1980 Int. Opt. Comput. Conf.*, SPIE, vol. 231, 1980, pp. 84-93.

[6] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, pp. 2758-2769, Aug. 1, 1982.

[7] J. R. Fienup, T. R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation," *J. Opt. Soc. Amer.*, vol. 72, pp. 610-624, May 1982.

[8] Yu. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," *Opt. Commun.*, vol. 30, pp. 304-308, Sept. 1979.

[9] G. H. Stout and L. H. Jensen, *X-Ray Structure Determination*. New York: Macmillan, 1968.

[10] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237-246, 1972.

[11] W. O. Saxton, *Computer Techniques for Image Processing in Electron Microscopy*. New York: Academic, 1978.

Appendix F

AMBIGUITY OF PHASE RETRIEVAL USING BOUNDARY CONDITIONS

J.R. Fienup

Submitted for publication in the Journal of the Optical Society of America.

ERIM

AMBIGUITY OF PHASE RETRIEVAL USING BOUNDARY CONDITIONS

J.R. Fienup

Environmental Research Institute of Michigan
P.O. Box 8618
Ann Arbor, Michigan 48107

ABSTRACT

It is shown that knowledge of the edges of an object is not always sufficient to uniquely reconstruct an object from the modulus of its Fourier transform via the autocorrelation function. On the other hand, in some cases not only can the boundary values be determined from the autocorrelation, but also the object can be reconstructed uniquely, even for complex-valued objects.

1. Introduction

In a number of disciplines, including astronomy, x-ray crystallography, electron microscopy, and wavefront sensing, one encounters the phase retrieval problem. One wishes to reconstruct $f(m, n)$, an object function, from $|F(p, q)|$, the modulus of its Fourier transform, where

$$F(p, q) = |F(p, q)| \exp [i\psi(p, q)] = \mathcal{F}[f(m, n)]$$

$$= \sum_{m=0}^{P-1} \sum_{n=0}^{Q-1} f(m, n) \exp [-i2\pi(mp/P + nq/Q)], \qquad (1)$$

where $m, p = 0, 1, \ldots, P - 1$ and $n, q = 0, 1, \ldots, Q - 1$. The discrete transform is employed here since in practice one deals with sampled data in a computer. The problem of reconstructing the object from its Fourier modulus is equivalent to reconstructing the Fourier phase, $\psi(p, q)$, from the Fourier modulus, since once one has the phase as well as the modulus, one can easily compute $f(m, n)$ by the inverse (discrete) Fourier transform. $r_f(m, n)$, the (aperiodic) autocorrelation of $f(m, n)$, is given by[1]

$$r_f(m, n) = \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} f(j, k) f^*(j - m, k - n) \qquad (2)$$

$$= \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} f^*(j, k) f(j + m, k + n) \qquad (3)$$

$$= \mathcal{F}^{-1}[|F(p, q)|^2]. \quad (4)$$

where the asterisk denotes complex conjugate and where it is assumed that $f(j, k) = 0$ for m outside of $[0, M - 1]$ and for n outside of $[0, N - 1]$. Note that in order to avoid aliasing in the computation of $|F(p, q)|^2$ it is necessary to have $M \leq P/2$ and $N \leq Q/2$. Since the autocorrelation function is easily computed from the Fourier modulus by Eq. (4), the phase retrieval problem is equivalent to reconstructing an object from its autocorrelation function.

Several phase retrieval algorithms have been proposed, all of them requiring some additional measurements or constraints on the solution. Examples include a reference point at least one object-diameter from the object[2] (giving rise to the holography condition[3]), a second intensity measurement in another plane[4-5] (in electron microscopy or wavefront sensing), nonnegativity and limited spatial extent[6-8] (in astronomy), atomic models[9] (in x-ray crystallography), objects consisting of collections of points having nonredundant spacings[10], and objects having latent reference points[11] (not satisfying the holography condition). For each of these situations there is a proof of uniqueness of the solution that relies on the types of measurements made, on the a priori information available, or on the nature of the reconstruction algorithm itself.

Antoher proposed phase retrieval algorithm is a recursive one that relies on a priori knowledge of the boundary conditions (i.e. the values of the edges of the object).[12] The purpose of this paper is to show that the general uniqueness claims made concerning phase retrieval using boundary conditions[12] are incorrect; but by the approach of using latent reference points,[11] special classes of objects can be shown to be unique, for complex-valued objects as well as for real-valued objects.

2. Ambiguity Using Boundary Conditions

In Reference 12 a recursive algorithm was put forward for reconstructing an object from the modulus of its Fourier transform, via the autocorrelation function, using boundary conditions, i.e., assuming knowledge of the edges of the object. A real-valued object, $f(m, n)$, was assumed to be zero outside of the rectangular region of support $0 \leq m \leq M - 1$ and $0 \leq n \leq N - 1$. The top and bottom nonzero rows, $\beta(m) = f(m, N - 1)$ and $\alpha(m) = f(m, 0)$, respectively, and the leftmost and rightmost nonzero columns, $f(0, n)$ and $f(M - 1, n)$, respectively, are assumed to be known a priori. Rows 1 and $N - 2$ can then be determined by solving a system of $2M - 1$ linear equations in $2M - 4$ unknowns. For example, from Eq. (3) we have, for $n = N - 2$, the second from the top row of the autocorrelation:

$$r(m, N - 2) = \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} f^*(j, k) f(j + m, k + N - 2)$$

$$= \sum_{j=0}^{M-1} f^*(j, 0) f(j + m, N - 2) + \sum_{j=0}^{M-1} f^*(j, 1) f(j + m, N - 1)$$

$$= \sum_{j=0}^{M-1} \alpha^*(j) f(j + m, N - 2) + \sum_{j=0}^{M-1} f^*(j, 1) \beta(j + m) \qquad (5)$$

for $m = -M + 1, \ldots, M - 1$. These are $2M - 1$ equations, one for each value of m, in $2M - 4$ unknowns, $f(j, N - 2)$ and $f(j, 1)$, for $j = 1, 2, \ldots, M - 2$. Recall that $\alpha(j)$, $\beta(j)$, $f(0, N - 2)$, $f(M - 1, N - 2)$, $f(0, 1)$ and $f(M - 1, 1)$ are assumed known. After $f(j, N - 2)$ and $f(j, 1)$ are determined by solving the system of equations given in Eq. (5) above, then one can solve for $f(j, N - 3)$ and $f(j, 2)$ using $r(m, N - 3)$ in a similar manner. The remaining rows of the object are solved

recursively in a similar manner.

The method described above could work if the systems of equations have a unique solution for the unknowns. Restricting the solution to real-valued f's, a claim has been made that,[12] "It may be shown, however, that a sufficient condition for a unique solution ... to exist is that $\alpha(m)$ and $\beta(m)$ not be identically zero and that $\alpha(m)$ not be related to $\beta(M - 1 - m)$ by a constant scale factor." However, no proof of that statement was provided. A counterexample to that claim is shown in Figure 1. Figures 1(a) and 1(b) show two different functions having the same boundaries as each other, and for both objects $\alpha(m)$ is not proportional to $\beta(M - 1 - m)$, and yet they have the same Fourier modulus and the same autocorrelation function, which is shown in Figure 1(c). Therefore knowledge of the boundaries is not necessarily sufficient information for a unique reconstruction.

An infinite number of counterexamples can be generated. From the theory of Bruck and Sodin[13] it is known that the solution of the phase retrieval problem [but not necessarily of Eq. (5)] is unique unless the Fourier transform of the object is a factorable polynomial, which is unlikely to happen by chance for the two-dimensional case. Factorability of the Fourier transform is equivalent to the object being expressible as a convolution of two functions, and so ambiguous cases can be constructed by forming an object by convolving (or cross-correlating) two functions.[14] The object in Figure 1(a) was fabricated by cross-correlating the functions shown in Figures 2(a) and 2(b). The ambiguous solution shown in Figure 1(b) is the inverted convolution of the functions shown in Figures 2(a) and 2(b). An infinite number of other ambiguous examples can be obtained by replacing the values 1, 1, 1 and 2 of the function shown in Figure 2(b) by other values, with minor restrictions on those values.

The recursive algorithm[12] involves the solution of 2M - 1 linear equations in 2M - 4 unknowns. One problem with this is that for m = -M + 1 and for m = M - 1, Eq. (5) involves only the known boundary values and not the unknowns. Therefore one has only 2M - 3 linear equations in 2M - 4 unknowns to begin with. A second problem is that upon inspection of those equations one finds that, for the ambiguous case shown in Figure 1, two or more of them are dependent equations. Since the number of remaining linear independent equations is fewer than the number of unknowns, the problem is underdetermined and multiple solutions exist.

Consider the particular example of Figure 1(c), for which one searches for solutions of the form shown in Figure 2(c), having the a priori known boundary values. Of the 2M - 3 = 7 linear equations of Eq. (5), utilizing the second row of Figure 1(c), one finds that three are dependent, leaving only four independent equations in six unknowns. Therefore one can, for example, choose values a and b in Figure 2(c) arbitrarily, and then the values of c, d, e and f are determined. At this point the algorithm of Reference 12 would have been stopped. However, if one continues to solve the equations using the next row of the autocorrelation, then one arrives at a quadratic equation in one of the variables yielding exactly two solutions, those shown in Figures 1(a) and 1(b).

From the example discussed above it is seen that the recursive algorithm of Reference 12 is much like the recursive algorithm of Dallas[5], in which a tree of solutions may grow with each iteration, and ambiguities are resolved only if the tree can be pruned in later iterations.

3. Some Unique Cases

Despite the nonuniqueness demonstrated in the previous section, there are some specific classes of objects for which the solution is unique. These unique objects have supports (or shapes) of special types.

Certain classes of objects having latent reference points can be reconstructed using a simpler recursive algorithm than the one described in the previous section. The simpler recursive algorithm[11] selects the order of the equations being solved such that at each step one must solve only a single linear equation for a single unknown, which is a trivial computation that always gives a unique result. It is required that no division by zero be allowed and this is ensured by the requirement that the values of the latent reference points not be zero. The latent reference points act in a similar manner to reference points for holography, only they do not initially satisfy the holographic separation condition. Examples of objects that can be uniquely reconstructed in this manner include (Fiddy-Brames-Dainty[15]) objects within a rectangle plus a point off one corner of the rectangle, and objects having other supports as well.[11] In most cases the support of the object must be known a priori in order to ensure that one obtains a unique reconstruction, since it is usually not possible to deduce the support of the object from the support of its autocorrelation[10]. However, for the Fiddy-Brames-Dainty[15] objects the support can be deduced from the autocorrelation support, and so the reconstruction in that case is unconditionally unique.[11] For these cases the objects may be complex-valued, in contrast with the restriction to real-valued objects for the reconstruction algorithm discussed in the previous section. Furthermore, for these cases the boundary values need not be known a priori since they are computed in the first step of the recursive algorithm[11,12].

4. Conclusions

Although boundary conditions are a powerful constraint for the phase retrieval problem, it has been proven by counterexample that knowledge of the boundary conditions (the values of the edges of the object) is not sufficient to ensure a unique solution. In practice it may be that a unique solution is usually obtained simply because 2-D phase retrieval is usually unique even when the boundary conditions are not known.[13] It is not yet known what extra constraints are necessary to ensure uniqueness in general.

What seems to be more important to ensure uniqueness is that the object's support be a member of a special class of supports. It is not yet known in general exactly what properties the support must have (except for the special cases mentioned in the previous section) to ensure uniqueness; but it is known that objects with separated supports[16,17] are more likely to be unique (even in the 1-D case) and objects having complicated supports tend to be easier to reconstruct than objects with convex symmetric support in the 2-D case[18].

The value of the recursive algorithms may be more in their predictions of uniqueness than in their ability to reconstruct images, since they tend to be very sensitive to noise[11,12]. A more stable reconstruction method would be the iterative Fourier transform approach[6], which repeatedly reinforces both the measured data and the a priori constraints on the reconstructed image.

Acknowledgement

This research was supported by the U.S. Air Force Office of Scientific Research under Contract F49620-82-K-0018.

REFERENCES

1. R.N. Bracewell, The Fourier Transform and Its Applications, 2nd Edition (McGraw-Hill, New York, 1978).

2. C.Y.C. Liu and A.W. Lohmann, "High Resolution Image Formation through the Turbulent Atmosphere," Opt. Commun. 8, 372-377 (1973).

3. J.W. Goodman, "Analogy Between Holography and Interferometric Image Formation," J. Opt. Soc. Am. 60, 506-509 (1970).

4. R.W. Gerchberg and W.O. Saxton, "A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures," Optik 35, 237-246 (1972); W.O. Saxton, Computer Techniques for Image Processing in Electron Microscopy (Academic Press, New York, 1978); R.H. Boucher, "Convergence of Algorithms for Phase Retrieval from Two Intensity Distributions," in 1980 International Optical Computing Conference, Proc. SPIE 231, 130-141 (1980).

5. W.J. Dallas, "Digital Computation of Image Complex Amplitude from Image- and Diffraction-Intensity: An Alternative to Holography," Optik 44, 45-59 (1975).

6. J.R. Fienup, "Phase Retrieval Algorithms: A Comparison," Appl. Opt. 21, 2758-2769 (1982).

7. J.R. Fienup, "Reconstruction of an Object from the Modulus of Its Fourier Transform," Opt. Lett. 3, 27-29 (1978); J.R. Fienup, "Space Object Imaging Through the Turbulent Atmosphere," Opt. Eng. 18, 529-534 (1979).

8. P.J. Napier and R.H.T. Bates, "Inferring Phase Information from Modulus Information in Two-Dimensional Aperture Synthesis," Astron. Astrophys. Suppl. 15, 427-430 (1974).

9. G.H. Stout and L.H. Jensen, X-Ray Structure Determination (Macmillan, London, 1968).

10. J.R. Fienup, T.R. Crimmins, and W. Holsztynski, "Reconstruction of the Support of an Object from the Support of Its Autocorrelation," J. Opt. Soc. Am. 72, 610-624 (1982).

11. J.R. Fienup, "Reconstruction of Objects Having Latent Reference Points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

12. M.H. Hayes and T.F. Quatieri, "Recursive Phase Retrieval Using Boundary Conditions," J. Opt. Soc. Am. 73, 1427-1433 (1983).

13. Yu.M. Bruck and L.G. Sodin, "On the Ambiguity of the Image Reconstruction Problem," Opt. Commun. 30, 304-308 (1979).

14. J.R. Fienup, "Image Reconstruction for Stellar Interferometery," in Current Trends in Optics, ed. F.T. Arecchi and F.R. Aussenegg (Taylor and Francis, London, 1981), pp. 95-102.

15. M.A. Fiddy, B.J. Brames and J.C. Dainty, "Enforcing Irreducibility for Phase Retrieval in Two Dimensions," Opt. Lett. 8, 96-98 (1983).

16. A.H. Greenaway, "Proposal for Phase Recovery from a Single Intensity Distribution," Opt. Lett. 1, 10-12 (1977); T.R. Crimmins and J.R. Fienup, "Ambiguity of Phase Retrieval for Functions with Disconnected Support," J. Opt. Soc. Am. 71, 1026-1028 (1981).

17. T.R. Crimmins and J.R. Fienup, "Uniqueness of Phase Retrieval for Functions with Sufficiently Disconnected Support," J. Opt. Soc. Am. 73, 218-221 (1983).

18. J.R. Fienup, "Experimental Evidence of the Uniqueness of Phase Retrieval from Intensity Data," in URSI/IAU Symposium on Indirect Imaging Conference Proceedings (Cambridge University Press, in printing), Sydney, Australia, 30 Aug.-1 Sept., 1983.

FIGURE CAPTIONS

1. Counterexample to uniqueness claims[12]. Two different objects (a) and (b) have the same boundary values and also have the same Fourier modulus and the same autocorrelation (c).

2. Functions (a) and (b) which generate the object shown in Figure 1(a) by cross-correlation and in Figure 1(b) by convolution. The general form (c) of the objects which have the autocorrelation shown in Figure 1(c).

2	3	3	3	1
3	7	4	4	2
3	6	5	4	2
1	2	2	2	1

(a)

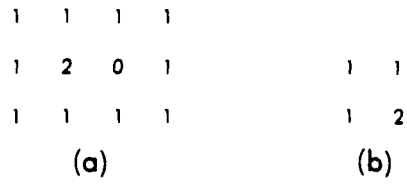
2	3	3	3	1
3	3	6	6	2
3	4	5	6	2
1	2	2	2	1

(b)

2	7	13	19	21	17	11	5	1
7	27	52	77	88	73	48	23	5
13	52	100	153	179	147	97	48	11
15	63	123	188	230	188	123	63	15
11	48	97	147	179	153	100	52	13
5	23	48	73	88	77	52	27	7
1	5	11	17	21	19	13	7	2

(c)

Figure 1. Counterexample to uniqueness claims.[12] Two different objects, (a) and (b), have the same boundary values and also have the same Fourier modulus and the same autocorrelation (c).



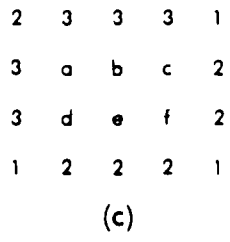


Figure 2. Functions (a) and (b) which generate the object shown in Fig. 1(a) by cross-correlation and in Fig. 1(b) by convolution. The general form (c) of the objects which have the autocorrelation shown in Fig. 1(c).

END

FILMED

2-85

DTIC